A Discussion Game for the Credulous Decision Problem of Abstract Dialectical Frameworks under Preferred Semantics

Keshavarzi Zafarghandi, Atefeh

Published in:
Online Handbook ofArgumentation for AI

Online Handbook of Argumentation for AI

Volume 1

Edited by

Federico Castagna
Francesca Mosca
Jack Mumford
Ștefan Sarkadi
Andreas Xydis

June 2020

Preface

This volume contains revised versions of the papers selected for the first volume of the Online Handbook of Argumentation for AI (OHAAI). Previously, formal theories of argument and argument interaction have been proposed and studied, and this has led to the more recent study of computational models of argument. Argumentation, as a field within artificial intelligence (AI), is highly relevant for researchers interested in symbolic representations of knowledge and defeasible reasoning. The purpose of this handbook is to provide an open access and curated anthology for the argumentation research community. OHAAI is designed to serve as a research hub to keep track of the latest and upcoming PhD-driven research on the theory and application of argumentation in all areas related to AI. The handbook's goals are to:

1. Encourage collaboration and knowledge discovery between members of the argumentation community.
2. Provide a platform for PhD students to have their work published in a citable peer-reviewed venue.
3. Present an indication of the scope and quality of PhD research involving argumentation for AI.

The papers in this volume are those selected for inclusion in OHAAI Vol.1 following a back-and-forth peer-review process undertaken by the editors of OHAAI Vol.1. The volume thus presents a strong representation of the current state of the art research of argumentation in AI that has been strictly undertaken during PhD studies. Papers in this volume are listed alphabetically by author. We hope that you will enjoy reading this handbook.

Editors
Federico Castagna
Francesca Mosca
Jack Mumford
Ștefan Sarkadi
Andreas Xydis

June 2020

Acknowledgements

We thank the senior researchers in the area of Argumentation and Artificial Intelligence for their efforts in spreading the word about the OHAAI project with early-career researchers.

We are also grateful to ArXiv for their willingness to publish this handbook.

We are especially thankful to Costanza Hardouin for designing the OHAAI logo.

We owe many thanks to Sanjay Modgil for helping to form the motivation for the handbook, and to Elizabeth Black and Simon Parsons for their advice and guidance that enabled the OHAAI project to come to fruition.

We owe special thanks to the contributing authors: Federico Castagna, Timotheus Kampik, Atefeh Keshavarzi Zafarghandi, Mickaël Lafages, Jack Mumford, Christos T. Rodosthenous, Samy Sá, Ștefan Sarkadi, Joseph Singleton, Kenneth Skiba, Andreas Xydis. Thank you for making the world of argumentation greater!

Contents

Argument Games for Dialectical Classical Logic Argumentation

Federico Castagna

Department of Informatics, King's College London, UK

Abstract

Argument games proof theories allow computing the membership of an argument to a specific extension according to the semantics the proof theory is meant to capture. These games assume the form of a dialectical exchange of arguments between two players which, alternating in turns, try to attack each other counterpart's arguments. Dialectical Classical logic Argumentation (Dialectical Cl-Arg) is a novel approach that provides real-world dialectical characterisations of Cl-Arg arguments by resource-bounded agents while preserving the rational criteria established by the rationality postulates. This paper combines both subjects and introduces argument games for Dialectical Cl-Arg, highlighting the properties and strengths enjoyed by these games in comparison with the standard ones. The resulting proof theory will better approximate real-world non-monotonic single-agent reasoning processes, bridging in this way the gap existing between formal and informal reasoning.

1 Introduction

Human reasoning evolved to produce and evaluate arguments ([Mercier and Sperber, 2011]). Trying to consolidate possessed information by formulating reasons (arguments) that challenge or defend the information itself, is an everyday procedure in which humans engage. This process is not only common but even necessary: how could be possible, otherwise, to decide what to believe or trust without being misled by a non-reliable source of information? This 'scaffolding' (as defined in [Modgil, 2017]) role of dialogue and arguments can also be seen in lone thinking practices since the reasoner will evaluate the possessed information by constructing counter-arguments against it and by assessing its reliability. That is to say, every reasoning process entails dialogue (even if it is just an imaginary dialogue that a person makes 'within himself/herself') and every dialogue entails arguments. The outlined reasoning process can be adapted for any type of agent-to-agent interaction: humans and artificial intelligences (henceforth AIs), among themselves and with humans. Thanks to its important role, argumentation has been developed as a theory able to characterize the essence of non-monotonic reasoning through the dialectical interplay of arguments [Dung, 1995]. Intuitively, in order to determine if a piece of information is reliable, it will suffice to show that the argument (in which the specific information is embedded) is justified under one of Dung's semantics. A way of doing this is to show the membership of the argument to a winning strategy of an argument game as described, for examples, in [Modgil and Caminada, 2009], [Vreeswik and Prakken, 2000] and [Caminada and Wu, 2009].

Although a plethora of works has successfully shown instantiations of Dung's abstract argumentation framework (AF) and reached different goals, none of these approaches managed to closely approximate the spontaneity of an everyday real-world interplay of arguments. With the introduction of the rationality postulates ([Caminada and Amgoud, 2007] and

[Caminada et al., 2012]), some steps have been moved in this direction by, for example, avoiding the arising of counterintuitive results in AF instantiations. However, this is still not enough. If we want to bridge the gap existing between formal and informal reasoning, we need to properly account for real-world uses of arguments by resource-bounded agents.

Stemming from a novel approach that provides real-world dialectical characterisations of AF by resource-bounded agents while preserving the rationality postulates ([D'Agostino and Modgil, 2018]), this paper will give a short description of its proof theory. The resulting dialectical argument game (fully-fledged developed as part of my PhD research) will better approximate a real-world non-monotonic single-agent reasoning process. That is to say, the inner process that an agent will go through in order to justify a piece of information it posses.

2 Method

This research made use of *(a)* the proof-theoretical method presented in [Modgil and Caminada, 2009] in order to develop an argument game for *(b)* Dialectical Cl-Arg [D'Agostino and Modgil, 2018].

(a) This method describes the general structure, the legal moves allowed and the winning conditions of a standard argument game. The precise protocol depends on the semantics which the proof theory is meant to capture. In a nutshell, an argument game is played by two players: a proponent (PRO) and its opponent (OPP). The proponent starts by moving an argument that it wants to test, after which each player must attack the other player's arguments with a counter-argument of sufficient strength. PRO wins the game if it is able to successfully defend against any counter-arguments moved by OPP. It loses otherwise.

(b) Dialectical Cl-Arg builds a formalization, for classical logic argumentation[1], that considers real-world dialectical exchange of arguments by resource-bounded agents. This entails:

- A new internal structure of the arguments is employed.
- Due to the limited availability of resources to real-world agents, only a finite subset of all the arguments of the AF will be taken into account (namely, $pdAF$), while still preserving satisfaction of the rationality postulates;
- The subset minimality and the consistency check on premises, required by Cl-Arg, are computationally unfeasible for resource-bounded agents. This is why the properties of Dialectical Cl-Arg allow to avoid them, while still preserving satisfaction of the rationality postulates.

3 Discussion

Argument games for Dialectical Cl-Arg are represented as trees branching downwards (called dialectical dispute trees). The roots of these trees correspond to the argument X that the proponent wants to test. If PRO is capable of defending X against any defeats[2] moved by the opponent's arguments and OPP runs out of legal moves according to the protocol of the specific game played (this depends on which Dung's semantics is considered), then PRO wins the game. The victory of the proponent implies that the piece of information embedded in X is reliable and justified according to the semantics the game was meant to capture.

To develop this proof theory, we had to adapt the work of [Modgil and Caminada, 2009] keeping in mind all the unique features of Dialectical Cl-Arg. The most problematic of which is certainly the different structure of the arguments, since it includes

[1]Detailed descriptions of Classical Logic Argumentation can be found in [Besnard and Hunter, 2008] and [Gorogiannis and Hunter, 2011].

[2]Notice that the considered argumentation framework is based on the defeat relation among arguments rather than the attack relation.

suppositions. To clarify, assume that $X = (\Delta, \Gamma, \alpha)$ is a Dialectical Cl-Arg argument, while $X' = (\Delta, \alpha)$ is a Cl-Arg argument. Δ and α are called, respectively, premises and conclusion, while Γ represents the suppositions. In real-world dialectical interactions, it is a common practice to suppose the premises of the opponent's arguments (without committing to them[3]), in order to show inconsistencies or draw new conclusions. As an example, let us consider the following fictitious exchange of arguments happening between a prosecutor (which assumes the role of PRO) and a suspect in a courthouse, namely Mr Corleone (which assumes the role of OPP):

> **PROSEC** *"We have noticed the transfer of about 20 million dollars to your bank account on the incriminated day. Mr Corleone, we strongly suspect you have not licitly earned that money."*
>
> **MR-COR** *"Aunt Mary passed away about a month ago. She was a very wealthy and generous woman. The money I received was just the inheritance I was legally entitled to."*
>
> **PROSEC** *"Ok. Let's suppose, as you are saying, that your old aunt passed away a month ago. According to the documents we retrieved, we know that Mary Corleone died three years ago. This seems to contradict your story."*
>
> **MR-COR** *"I am talking about a different relative: Mary Rossi and not Mary Corleone."*
>
> **PROSEC** *"Very well. Let's then suppose that Mary Rossi is the relative from whom you received the money. However, no official record seems to certify the existence of this woman. This thwarts your story once again."*

The prosecutor supposes the premises of the suspect's arguments (hence, accepts the premises without committing to them) in order to derive conclusions that defeat the arguments moved by Mr Corleone, undermining the credibility of his defence.

In general, when challenging the acceptability of an argument with respect to an admissible set $\mathcal{S}$, the defeating argument can suppose premises from all the arguments in $\mathcal{S}$. Whereas, the argument that defends $\mathcal{S}$ can only suppose the premises of the defeating argument. Accommodating this in a dialectical dispute tree means that, when testing the acceptability of the arguments that PRO has moved in the winning strategy (i.e., the set $\mathcal{S}$), OPP can suppose the premises of these arguments in order to draw conclusions. PRO, in turn, can only suppose the premises of the argument that OPP is playing for invalidating the winning strategy. Although adding suppositions and the reference to a set $\mathcal{S}$ complicates the formalisms of the argument games, it also enables additional dialectical moves to the players, better approximating a real-world reasoning process. An instance of the newly introduced proof theory can be seen in Figure 1. Starting with the

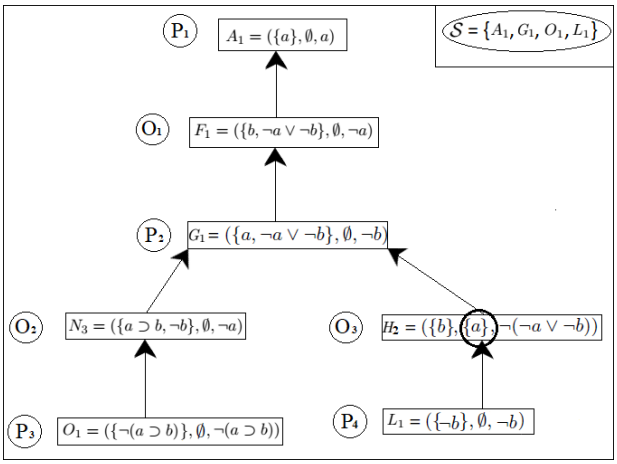


Figure 1: An example of a dialectical dispute tree

root A_1, which is the argument that PRO wants to test, the two players alternate in moving arguments extending the dialectical dispute tree following the order highlighted by the numbers near the labels P and O (meaning, respectively, PRO and OPP). Argument H_2, played by OPP, supposes the premise a (circled in Figure 1) of either A_1 or G_1, since both of them are in $\mathcal{S}$. This supposition allows H_2 to derive a conclusion, which defeats a premise of G_1[4]. However,

[3] This allows avoiding the so-called "Foreign commitment problem" [Caminada et al., 2014].

[4] Notice that the only allowed defeat is the undermine defeat

the proponent succeeded in defending the root A_1 against each of the opponent's defeat. Assuming OPP has no more legal moves to play, PRO has a winning strategy and wins the game. This implies that A_1 is an argument justified according to the semantics of the played game, i.e., the information embedded by A_1 is reliable according to the semantics the game was meant to capture.

Dialectical dispute tree properties

In the following, we are going to list the main features enjoyed by any dialectical dispute tree in comparison with the standard (non-dialectical) dispute trees. We will not consider most of the properties inherited from Dialectical Cl-Arg since they are not useful for this purpose:

> Set $\mathcal{S}$ *The arguments moved by the proponent in the winning strategy correspond to the admissible set $\mathcal{S}$, which includes and defends the root of the tree. That is to say, the information that PRO wants to test is defended and made reliable by the arguments in $\mathcal{S}$.*
>
> Relevance *The players alternate to move arguments that change the outcome of the game at every turn, avoiding any unnecessary detour to this task.*
>
> Minimality of the winning strategy *Real-world agents do not waste their limited amount of resources. PRO moves only the minimal number of arguments needed for generating a winning strategy.*
>
> No conflicting PRO arguments *The detection of conflicting arguments in $\mathcal{S}$ happens via dialectical means. This prevents PRO from building a winning strategy which is not conflict-free.*
>
> No self-defeating arguments *No rational real-world agent would state an argument that defeats itself. This move would be useless for the proceeding of the game and, as such, it would be a misuse of resources.*

The above properties outline a dispute tree composed by moves more aligned with the real-world uses of arguments for resource-bounded agents.

4 Conclusion

The main features of the real-world uses of argumentation by resource-bounded agents include: (a) showing arguments inconsistencies by supposing the opponent's premises; (b) handling only finite subsets of the arguments of the AFs; (c) optimizing resources consumption by employing dialectical means (while still satisfying the rationality postulates). These attributes constitute the main components of the introduced argument game proof theory, thus capable of better approximate non-monotonic single-agent reasoning processes. Unfortunately, the limited space prevented us to fully appreciate the extent of the formalism which would have also included specific protocols for Dung's grounded and preferred semantics.

The overall aim of my PhD is to investigate and develop proof theories for Dialectical Cl-Arg. As such, a natural extension of the research presented in this paper will be the generation of algorithms for computing argument extensions through an adaptation of the method of labelling described in [Modgil and Caminada, 2009]. The labelling approach has the advantage of easily bringing the dialectical reasoning of the argument games to an algorithmic level. This work will then be further expanded to include argument games and labellings for the stable ([Caminada and Wu, 2009]), semi-stable ([Caminada, 2007]) and ideal semantics ([Dung et al., 2007] [Caminada, 2011]). If time permits, another research path that will be pursued will involve the generalisation of the developed dialectical argument games to dialogues by following the guidelines of the already existing literature in

(i.e., the defeat that targets the premises of an argument) [D'Agostino and Modgil, 2018]. Also, for simplicity, we are omitting the preference relation existing among the arguments of the dialectical dispute tree.

the field (mainly [Prakken, 2005]). This would have the interesting consequence of allowing to move from non-monotonic single-agent inference to distributed non-monotonic reasoning.

Acknowledgements

I would like to thank S.Modgil and M.D'Agostino for their precious help and valuable suggestions: without them, this research could not have been possible.

References

[Besnard and Hunter, 2008] Besnard, P. and Hunter, A. (2008). *Elements of argumentation*, volume 47. MIT press Cambridge.

[Caminada, 2007] Caminada, M. (2007). An algorithm for computing semi-stable semantics. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 222–234. Springer.

[Caminada, 2011] Caminada, M. (2011). A labelling approach for ideal and stage semantics. *Argument and Computation*, 2(1):1–21.

[Caminada and Amgoud, 2007] Caminada, M. and Amgoud, L. (2007). On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5-6):286–310.

[Caminada et al., 2012] Caminada, M., Carnielli, W., and Dunne, P. (2012). Semi-stable semantics. *Journal of Logic and Computation*, 22(5):1207–1254.

[Caminada et al., 2014] Caminada, M., Modgil, S., and Oren, N. (2014). Preferences and unrestricted rebut. *Computational Models of Argument*.

[Caminada and Wu, 2009] Caminada, M. and Wu, Y. (2009). An argument game for stable semantics. *Logic Journal of IGPL*, 17(1):77–90.

[D'Agostino and Modgil, 2018] D'Agostino, M. and Modgil, S. (2018). Classical logic, argument and dialectic. *Artificial Intelligence*, 262:15–51.

[Dung, 1995] Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.

[Dung et al., 2007] Dung, P. M., Mancarella, P., and Toni, F. (2007). Computing ideal sceptical argumentation. *Artificial Intelligence*, 171(10-15):642–674.

[Gorogiannis and Hunter, 2011] Gorogiannis, N. and Hunter, A. (2011). Instantiating abstract argumentation with classical logic arguments: Postulates and properties. *Artificial Intelligence*, 175(9-10):1479–1497.

[Mercier and Sperber, 2011] Mercier, H. and Sperber, D. (2011). Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74.

[Modgil, 2017] Modgil, S. (2017). Dialogical scaffolding for human and artificial agent reasoning. In *AIC*, pages 58–71.

[Modgil and Caminada, 2009] Modgil, S. and Caminada, M. (2009). Proof theories and algorithms for abstract argumentation frameworks. *Argumentation in artificial intelligence*, 105129.

[Prakken, 2005] Prakken, H. (2005). Coherence and flexibility in dialogue games for argumentation. *Journal of logic and computation*, 15(6):1009–1040.

[Vreeswik and Prakken, 2000] Vreeswik, G. A. and Prakken, H. (2000). Credulous and sceptical argument games for preferred semantics. In *European Workshop on Logics in Artificial Intelligence*, pages 239–253. Springer.

Economic Rationality and Abstract Argumentation

Timotheus Kampik

Umeå University, Sweden

Abstract

This article presents a line of work that builds a bridge between abstract argumentation as a method of non-monotonic reasoning and formal models of economically rational decision-making. As the foundation of this bridge, we introduce the *reference independence* principle, which is a key property of economic rationality, to abstract argumentation. We relate this principle to principles of non-monotonic reasoning and, from this starting point, outline a set of research directions we are pursuing to better integrate abstract argumentation and models of economic rationality.

1 Introduction

In the symbolic artificial intelligence community, formal argumentation has emerged as a popular approach to instill reasoning capabilities into intelligent systems. A foundational method of formal argumentation is *abstract argumentation* [Dung, 1995]. An abstract argumentation framework is a tuple of atomic arguments and binary relations (*attacks*) between these arguments. For example, when we construct the argument framework $AF = (\{a, b\}, \{(a, b)\})$, we have the arguments a and b, and argument b is attacked by argument a. Arguments can be, for example, epistemic (let a denote the fact that it rains) or utilitarian (let b denote the action of leaving the house without an umbrella). To determine which set(s) of arguments in an argumentation framework can be considered "feasible" conclusions, argumentation semantics have been defined. While it is clear that the set of conclusions that results from the argumentation framework AF is $\{a\}$, determining conclusions is not trivial for cyclic argumentation frameworks. For example, in the argumentation framework $AF' = (\{a, b, c\}, \{(a, b), (b, c), (c, a)\})$, either no arguments ($\{\}$), or any of the sets $\{a\}$, $\{b\}$, or $\{c\}$ can be considered conclusions (depending on the semantics). Consequently, an argumentation semantics can return several argument sets that can be potentially be considered acceptable for a given framework. In this article we use *preferred semantics* as defined in the initial paper on abstract argumentation [Dung, 1995]. Given an argumentation framework $AF = (AR, AT)$, let us first define a conflict-free set of arguments as a set $S \subseteq AR$ that do not attack each other. Also, a set $S \subseteq AR$ is admissible iff it is conflict-free and its arguments attack all arguments in AR that attack S. A set $S \subseteq AR$ is a preferred extension of AF iff it is maximal with regards to set inclusion among all admissible sets in AF. Preferred semantics determines the preferred *extensions* of AF. Let us denote all preferred extensions of AF by $\sigma_{pref}(AF)$. Given the two example frameworks above, we have $\sigma_{pref}(AF) = \{\{a\}\}$ and $\sigma_{pref}(AF') = \{\{\}\}$.

Many different argumentation semantics exist, and it is often not clear which semantics is the most feasible one for a specific application scenario. Consequently, an important line of research on abstract argumentation is the identification of *argumentation principles*–formal properties of argumentation semantics–and the evaluation of argumentation semantics w.r.t. to these

principles [van der Torre and Vesic, 2017]. In our line of research, we add a new perspective to the principle-based evaluation by introducing a principle that is based on microeconomic decision-theory. In particular, we introduce the *reference independence* principle, which serves as the point of departure for more research at the intersection of abstract argumentation and formal models of economic rationality. We derive this principle from the property of the same name that is a cornerstone of the *rational man* paradigm in microeconomic theory. When choosing items from a set S, a rational-decision maker's choice $A^* \subseteq S$ implies that the decision-maker prefers A^* over all other sets in 2^S. When choosing from another set that potentially intersects with S, the implied preferences must be consistent.

Example 1. *For instance, when we have a consumer who can choose to consume from a set containing tea and juice ($\{t, j\}$), her choice of juice ($\{j\}$) implies $\{j\}$ is preferred over all other sets in $2^{\{t,j\}}$. Let us assume that on another occasion, a third item–a donut–is present in the set, all other things being the same as before. Our consumer chooses tea and a donut ($\{t, d\}$ from $\{t, j, d\}$). The choice implies that $\{t, d\}$ is preferred over all other sets in $2^{\{t,j,d\}}$, which is consistent with the previous choice. However, were she to choose juice ($\{j\}$ from $\{t, j, d\}$), the preference $\{j\}$ over $\{t\}$ would be inconsistent with the previously established preference $\{t\}$ over $\{j\}$.*

Note that in our interpretation of the rational man model, the set of choice items does not necessarily need to refer to physical goods, but can also model courses of action, or beliefs that can be adopted or discarded. Our ambition to better integrate abstract argumentation and formal models of economic decision-making can be considered a natural continuation of work presented in the initial paper on abstract argumentation, which applies argumentation to the stable marriage problem of cooperative game theory [Dung, 1995].

Let us introduce and motivate the concept of reference independence in abstract argumentation with the help of a simple example.

Example 2. *Let us assume the role of a strategy advisor (human or IT system) in a large corporation. We propose the launch of new products to a decision-maker who has the final say. At the moment, we are considering the launch of the products a' or b'. a' and b' are similar; however, studies show that a' is expected to outperform b'. We model this assessment as an argumentation framework $AF = (AR, Attacks)$, such that:*

$$AF = (\{a, b\}, \{(a, b)\}),$$

where a means "launch a'" and b means "launch b'". Let us assume we resolve AF using preferred semantics σ_{pref}, i.e., *$\sigma_{pref}(AF) = \{\{a\}\}$. Consequently, we tell the decision-maker that she can launch a'.*

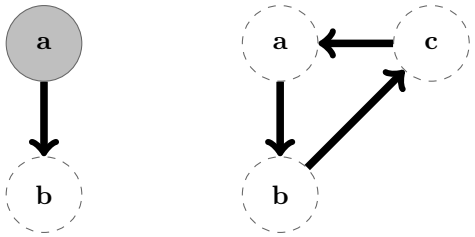
Now, let us assume the decision-maker postpones her decision and asks us to come back some time later with an updated analysis. In the meantime, a new product–c'–is prototyped and evaluated in terms of market fit by our R&D department. According to the evaluation, the target consumer group typically prefers buying c' over a', while they typically prefer b' over c':

$$AF' = (\{a, b, c\}, \{(a, b), (b, c), (c, a)\})$$

We again use preferred semantics, i.e., *$\sigma_{pref}(AF') = \{\{\}\}$. However, it is clear that the decision-maker will question our sanity if we recommend her to launch no product now that more potential options are on the table*[1]*. Indeed, the adjustment of our decision outcome from "a'" to "nothing" ($\{\}$) is inconsistent with the* reference independence *principle in microeconomic theory: the addition of the* irrelevant alternatives *$2^{AR'} \setminus 2^{AR}$ of argument sets we can potentially consider as acceptable makes us switch from accepting $\{a\}$ ($\{a\}$ is preferred over $\{\}$) to accepting $\{\}$ ($\{\}$ is preferred over $\{a\}$). Figure 1 depicts the example's argumentation frameworks.*

In this line of research we aim to address the problem the example highlights.

[1] A better recommendation would be to delay the decision until more intelligence is gathered. Still, it makes sense to be able to make a somewhat reasonable decision at any point.



(a) AF. (b) AF'.

Figure 1: Reference dependence: $\sigma_{pref}(AF) = \{\{a\}\}$ and $\sigma_{pref}(AF') = \{\{\}\}$. The addition of the not acceptable argument c makes us discard the extension $\{a\}$ in favor of $\{\}$.

2 Reference Independence and Cautious Monotony

The motivation of the reference independence principle is to assess whether a decision-making outcome can be considered "reasonable". In the domain of non-monotonic reasoning, the cautious monotony [Gabbay, 1985, Schröder et al., 2010] and rational monotony [Benferhat et al., 1997] principles have been defined with a similar objective, but without making the connection to economic decision theory. In our work, we introduce these principles to abstract argumentation. We define *strong* and *weak* restricted and rational monotony properties, similarly to the way Čyras and Toni have defined cautious monotony in the context of *assumption-based argumentation* [Čyras and Toni, 2015]. Let us have an argumentation semantics σ, two argumentation frameworks $AF = (AR, AT)$ and $AF' = (AR', AT')$, and their extensions $\sigma(AF)$ and $\sigma(AF')$. We can colloquially describe cautious and rational monotony as follows:

- For each extension E in $\sigma(AF)$, we "adjust" AF' and get an AF'' in which all "new" attacks (that are in AF', but not in AF) to E are removed. σ is **strongly cautiously monotonous** iff all extensions $E'' \in \sigma(AF'')$ contain E. σ is **weakly cautiously monotonous** iff there exists an extension $E'' \in \sigma(AF'')$ that contains E.
- For each extension E in $\sigma(AF)$, we "adjust" AF' and get an AF''' in which all "new" attacks to E, as well as from E, are removed. σ is **strongly rationally monotonous** iff all extensions $E''' \in \sigma(AF''')$ contain E. σ is **weakly rationally monotonous** iff there exists an extension $E''' \in \sigma(AF''')$ that contains E.

Analogously, we can describe reference independence as follows. Again, we have an argumentation semantics σ, two argumentation frameworks $AF = (AR, AT)$ and $AF' = (AR', AT')$, and their extensions $\sigma(AF)$ and $\sigma(AF')$:

- **Strong reference independence**. For each extension E in $\sigma(AF)$, the preferences over the argument sets in $2^{AR \cap AR'}$ implied by all extensions in $\sigma(AF')$ are consistent with the preferences implied by E.
- **Weak reference independence**. For each extension E in $\sigma(AF)$, the preferences over the argument sets in $2^{AR \cap AR'}$ implied by at least one extension in $\sigma(AF')$ are consistent with the preferences implied by E.

In [Kampik and Nieves, 2020], we provide a comprehensive formal comparison of reference independence and other non-monotonic reasoning properties in the context of abstract argumentation. We also show that most (but not all) argumentation semantics are not weakly reference independent. In ongoing research, we work on the definition of new semantics that are reference independent and also fulfill other desirable principles.

3 Ensuring Reference Independence

It is clear that strong reference independence is a property that is unrealistic to obtain. In contrast, weak reference independence can be considered useful, as we can show with the help of an example, in which we make use of concepts introduced by Gabbay for the purpose of *loop-busting* [Gabbay, 2014].

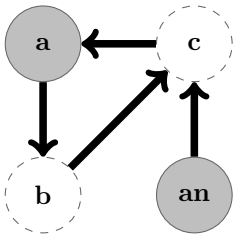


Figure 2: AF'_{an}. Solving the problem depicted in Figure 1. The annihilator approach enables us to achieve reference independence.

1. We go back to Example 2, starting with $AF = (\{a, b\}, \{(a, b)\}$, which we resolve using preferred semantics as $\sigma_{pref}(AF) = \{\{a\}\}$. Because we have exactly one extension, we decide that $\{a\}$ is the set of acceptable arguments (*i.e.*, we recommend launching product a').

2. When resolving the expanded framework $AF' = (\{a, b, c\}, \{(a, b), (b, c), (c, a)\})$, $\sigma_{pref}(AF')$ returns $\{\{\}\}$. This implies inconsistent preferences with regards to how we have resolved AF (as explained in Example 1). Hence, we create an argumentation framework AF'_{an}, in which an *annihilator argument* is added to AF', such that we have exactly one extension $E \in \sigma(AF'_{an})$ and $E \setminus \{an\}$ is an extension of AF' that implies consistent preferences with the extension $\{a\}$ we have previously determined for AF. For example, we can define AF'_{an} as $(\{a, b, c, an\}, \{(a, b), (b, c), (c, a), (an, c)\})$ so that $\sigma_{pref}(AF'_{an}) = \{\{a, an\}\}$. Given the only extension $E = \{a, an\}$, we have $E \setminus \{an\} = \{a\}$. $E \setminus \{an\}$ is our final extension of AF'.

3. For any subsequent argumentation framework, we can check if there is any extension that ensures reference independence with regards to the previous argumentation framework and, if not, proceed as described in the steps before.

Figure 2 depicts the argumentation AF'_{an}. In ongoing research, we work on defining formal approaches to ensure reference independence, as well as other non-monotonic reasoning principles, when resolving *sequences* of argumentation frameworks for semantics that do not fulfill these principles in general.

4 Game Theory and Abstract Argumentation

As mentioned above, already the initial paper on abstract argumentation relates to (cooperative) game theory. Further research provides game theoretical perspectives on abstract argumentation primarily by observing properties that emerge from the exchange of arguments between several autonomous agents. Thereby, no assumptions are made with regards to the agent's rationality in the formal economic sense. For instance, Rahwan and Larson show, for some argumentation semantics and depending on the properties of the agents' preferences, how the Pareto-optimal sets of arguments in an argumentation framework relate to the extensions different semantics return [Rahwan and Larson, 2008].

Given that we have introduced the formal foundations of instilling economically rational behavior into abstract argumentation-based agents, the existing works on argumentation and game theory can be examined from a different perspective. The results of this research direction can potentially be applied to define agreement protocols for autonomous agents.

5 Conclusion

In this article, we have provided an intuition of how principles of rational microeconomic decision-making can be applied to abstract argumentation. We have outlined a set of promising research directions to further advance research at the intersection of formal argumentation, non-monotonic reasoning and economic theory. We expect that the research results will shed new light on how abstract argumentation can be used as a non-monotonic reasoning method. Potentially, this line of work can enable the introduction of formal argumentation as a model

of economic decision-making to the microeconomics community.

Acknowledgements

The author thanks Dov Gabbay and Juan Carlos Nieves, who are collaborators and mentors for the presented line of research. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[Benferhat et al., 1997] Benferhat, S., Dubois, D., and Prade, H. (1997). Nonmonotonic reasoning, conditional objects and possibility theory. *Artif. Intell.*, 92(1–2):259–276.

[Čyras and Toni, 2015] Čyras, K. and Toni, F. (2015). Non-monotonic inference properties for assumption-based argumentation. In Black, E., Modgil, S., and Oren, N., editors, *Theory and Applications of Formal Argumentation*, pages 92–111, Cham. Springer International Publishing.

[Dung, 1995] Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321 – 357.

[Gabbay, 2014] Gabbay, D. (2014). The handling of loops in argumentation networks. *Journal of Logic and Computation*, 26(4):1065–1147.

[Gabbay, 1985] Gabbay, D. M. (1985). Theoretical foundations for non-monotonic reasoning in expert systems. In Apt, K. R., editor, *Logics and Models of Concurrent Systems*, pages 439–457, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Kampik and Nieves, 2020] Kampik, T. and Nieves, J. C. (2020). Abstract argumentation and the rational man.

[Rahwan and Larson, 2008] Rahwan, I. and Larson, K. (2008). Pareto optimality in abstract argumentation. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 1*, AAAI'08, page 150–155. AAAI Press.

[Schröder et al., 2010] Schröder, L., Pattinson, D., and Hausmann, D. (2010). Optimal tableaux for conditional logics with cautious monotonicity. In *Proceedings of the 2010 Conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, page 707–712, NLD. IOS Press.

[van der Torre and Vesic, 2017] van der Torre, L. and Vesic, S. (2017). The principle-based approach to abstract argumentation semantics. *IfCoLog Journal of Logics and Their Applications*, 4(8).

A Discussion Game for the Credulous Decision Problem of Abstract Dialectical Frameworks under Preferred Semantics

Atefeh Keshavarzi Zafarghandi

Department of Artificial Intelligence,
Bernoulli Institute, University of Groningen, The Netherlands

Abstract

Abstract dialectical frameworks (ADFs) have been introduced as a general formalism for modeling and evaluating argumentation. However, the role of discussion in reasoning in ADFs has not been clarified well so far. The current work presents a discussion game, as a proof method, to answer credulous decision problems of ADFs under preferred semantics. The game is the basis for an algorithm that can be used not only for answering the decision problem but also for human-machine interaction.

1 Introduction

Argumentation has recently received increased attention within artificial intelligence. A wide range of formalisms has been introduced for modeling and evaluating argumentation. *Abstract dialectical frameworks* (ADFs), introduced first by [Brewka and Woltran, 2010], are expressive generalizations of Dung's widely used argumentation frameworks (AFs) [Dung, 1995]. ADFs abstract away from the content of arguments but are expressive enough to model different types of relations among arguments. A key question is 'How is it possible to evaluate the truth value of arguments in a given ADF?' Answering this question leads to the introduction of several types of semantics, defined based on three-valued interpretations. Moreover, answering whether there exists an interpretation of a particular type of semantics in which an argument has a given value is a fundamental issue. In ADFs, an *admissible* interpretation does not contain any unjustifiable information about the arguments and *preferred semantics* are prominent semantics that present maximum information about the arguments without losing admissibility. Further, answering decision problems of preferred semantics has a higher computational complexity than other semantics in ADFs [Strass and Wallner, 2015]. Thus, answering them has a crucial importance.

Although dialectical methods have a role in determining semantics of both AFs and ADFs, the roles are not obvious in the definition of semantics. To cover this gap, quite a number of works have been presented to show that semantics of AFs can be interpreted in terms of structural discussion [Prakken and Sartor, 1997, Caminada, 2017, Dung and Thang, 2007]. Further, the presented methods have been used in human-machine interaction [Booth et al., 2018], which is a wide research area in AI.

Because of the special structure of ADFs, the existing methods used to interpret semantics of AFs cannot be reused in ADFs. To address this problem the first existing game for preferred semantics of ADFs is presented by [Keshavarzi Zafarghandi et al., 2019]. I am working on a modification of that game to reduce the computational complexity of the game in both best and average cases. The previous game is defined

based on only one type of move, named *forward move*. To reduce the complexity, in the current game the forward move is modified and a *backward move* is also defined. Moreover, based on the current method, an algorithm can be provided not only to answer credulous decision problems of ADFs under preferred semantics but also to be used in a human-machine dialogue. Suppose that an ADF is used to formalize a knowledge-base that presents methods to cure a disease. It is not enough to tell a patient that a chosen method is the best one because it is presented in a semantics, but the patient needs to be convinced why this is the case. The current work provides a discussion game as a proof method to cover this gap, for preferred semantics of ADFs. Further, the presented method is sound and complete. In Section 2 first I present a brief relevant background of ADFs and then I present the idea of the game.

2 Method

An *abstract dialectical framework* (ADF) is a tuple $F = (A, L, C)$ where: 1. A is a finite set of statements (arguments); 2. $L \subseteq A \times A$ is a set of links among arguments; 3. $C = \{\varphi_a\}_{a \in A}$ is a set of propositional formulas, called acceptance conditions [Brewka and Woltran, 2010]. Acceptance conditions indicate the set of links implicitly, thus, there is no need of presenting L in ADFs explicitly.

An *interpretation* v (for F) is a function $v : A \mapsto \{\mathbf{t}, \mathbf{f}, \mathbf{u}\}$ s.t. $\mathbf{t}$, $\mathbf{f}$ and $\mathbf{u}$ refer to true, false and undecided, respectively. Truth values can be ordered via the information ordering relation $<_i$ given by $\mathbf{u} <_i \mathbf{t}$ and $\mathbf{u} <_i \mathbf{f}$. Relation $\leq_i$ is the reflexive and transitive closure of $<_i$. Interpretations can be ordered via $\leq_i$ w.r.t. their information content.

Given an interpretation v, the partial valuation of φ_a by v, is $\varphi_a^v = \varphi_a[b/\top : v(b) = \mathbf{t}][b/\bot : v(b) = \mathbf{f}]$, s.t. b is a parent of a. Semantics for ADFs can be defined via the *characteristic operator* Γ_F. Applying Γ_F on v leads to v' s.t. for each $a \in A$, 1. $v'(a) = \mathbf{t}$ if φ_a^v is irrefutable (i.e. a tautology), 2. $v'(a) = \mathbf{f}$ if φ_a^v is unsatisfiable, 3. otherwise, $v'(a) = \mathbf{u}$.

An interpretation v is admissible if $v \leq_i \Gamma_F(v)$ and it is preferred if it is $\leq_i$-maximal admissible. It is said that a is credulously acceptable (deniable) under σ semantics, if there exists a σ-interpretation v for which φ_a^v is irrefutable (resp. unsatisfiable). Whenever there is no ambiguity, we write interpretations by the sequence of truth values, by choosing the lexicographic order on arguments. For instance, $v = \{a \mapsto \mathbf{t}, b \mapsto \mathbf{u}\}$ can be represented by $\mathbf{tu}$.

In a study, quite a number of women (71.5 percent) believed that mammography is a precise and safe method of diagnosing cancer. However, researchers have found that mammography has demonstrated a number of adverse effects, two of which are breast cancer over-diagnosis and causes of tumor rupture and spread of cancerous cells.[1] Using the ADF formalism, this knowledge base can be modeled by an ADF that contains three statements. Statement m : 'mammography is a precise and a safe method' is acceptable if and only if mammography neither causes o 'over-diagnosis' nor r 'rupture of cancer cells'. That is, the acceptance condition of m, namely φ_m can be represented by propositional formula $\neg o \wedge \neg r$. Since statements of o and r are facts, proven by recent research, they are always accepted. Thus, the acceptance conditions of them are $\varphi_r \equiv \top$ and $\varphi_o \equiv \top$.

Assume that a proponent (P) believes that mammography is a safe and precise method. To discuss about the belief, an opponent (O) checks the acceptance condition of m and says 'if P's belief is true, then by $\varphi_m : \neg o \wedge \neg r$ both o and r have to be denied.' Then, O challenges P: 'Do you have any reason why both o and r are deniable?' In the next step P checks the acceptance conditions of o and r and since both of them are tautologies, neither of them can be denied. Thus, the main belief of P is false. This corresponds with the fact that in this ADF, there is no preferred interpretation that satisfies the belief of P.

A *preferred discussion game* is a two-player game between proponent (P) and opponent (O), in which P presents a claim about credulous acceptance or denial of an argument under preferred semantics in a given ADF. A claim of P about the truth value of an argument can be represented by interpretation

[1] https://kresserinstitute.com/the-downside-of-mammograms/

v_0, called *initial claim*. In v_0 the argument which is claimed is assigned to **t** (resp. **f**) if it is claimed that it is accepted (denied). Since there is no further information about all other arguments, they are assigned to **u**. The ideas of the game, including forward and backward moves, are presented in Example 2.1.

Example 2.1. *Let* $F = (\{a, b\}, \{\varphi_a : \neg b, \varphi_b : \neg a \vee c, \varphi_c : \bot\})$ *be an ADF, depicted in Figure 1. Assume that the proponent claims that* b *is credulously acceptable under preferred semantics of* F*. The initial claim of* P *can be written by interpretation* $v_0 = \mathbf{utu}$.

- *The game is continued by* O *by applying a forward move that contains two steps. 1.* O *checks the consequence of* v_0 *on the the acceptance condition of the argument which is claimed by* P*. That is,* O *evaluates* φ_b *by* v_0*, which is* $\varphi_b^{v_0} : \neg a \vee c$*. Here* v_0 *does not have any role on satisfiability of* φ_b*. 2.* O *picks the truth value of the parents of* b *in* $\varphi_b^{v_0}$ *that can satisfy the claim. For instance,* O *says based on the acceptance condition of* b*,* b *can be accepted if* a *is denied. That is,* O *presents that 'I will agree with you on the truth value of* b *in a preferred interpretation if you can show that* $a \mapsto \mathbf{f}$ *in that interpretation'. This new piece of information can be presented by* $v_1 = \mathbf{ftu}$ *as the result of forward move. In other words,* O *challenges* P *by asking 'what is your reason of this assignment of* a*?'*

- *Since* $v_0 <_i v_1$*, the dialogue between players can be continued. Now it is* P*'s turn to investigate whether the challenge of* O *is satisfiable. First,* P *checks the role of* v_1 *on* φ_a*.* $\varphi_a^{v_1} : \bot$ *presents that* a *is deniable in a preferred interpretation in which* b *is acceptable. Thus, the forward move does not have the second step of finding the truth value of parents of* a *in* $\varphi_a^{v_1}$*. Therefore, the forward move of* P *leads to* $v_2 = \mathbf{ftu}$.

- *Since* $v_1 = v_2$*, the dialogue between the players stops.* $v_1 = v_2$ *means that the information of* v_1 *is enough to answer* O*'s challenge. Further,* P *answers the challenge of* O *without presenting a new claim. That is,* P *defends the initial claim. Thus, the game stops here and* P *wins the game.*

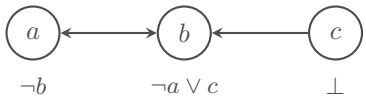


Figure 1: ADF of Example 2.1

On the other hand, if in a dialogue of a game $v_i \not\leq_i v_{i+1}$*, then the dialogue between players stops and* P *loses that dialogue. However, this does not mean that* P *loses the game. The reason of this is that possibly there exists another dialogue by which* P *can defeat a challenge of* O *to defend the initial claim. In this situation* P *applies a move, called* backward move *to find a new dialogue. The idea of this move is presented in the following. Consider that in ADF* F*, depicted in Figure 1,* O *presents the following challenge to challenge the initial claim.*

- O *says based on* $\varphi_b : \neg a \vee c$*,* b *can be accepted if* c *is acceptable. That is,* O *asks* P *'can you indicate whether* $c \mapsto \mathbf{t}$ *in a preferred interpretation in which* $b \mapsto \mathbf{t}$*?' Thus,* O*'s forward move is* $v_1 = \mathbf{utt}$.

- *Since* $v_0 <_i v_1$ *the dialogue between players continues. First,* P *has to check the consequence of the information presented by* v_1 *on the challenging argument, namely* c*. That is,* P *evaluates* $\varphi_c^{v_1} : \bot$*. Since* $\varphi_c^{v_1}$ *is unsatisfiable and* c *is assigned to* **t** *in move* v_1*,* P *cannot decide about the truth value of* c *in this move. Thus, the forward move of* P *leads to* $v_2 = \mathbf{utu}$.

- *Since* $v_1 \not\leq_i v_2$*, this dialogue cannot continue anymore. That is,* P *loses this dialogue, but not the game. That is,* P *may attempt to find a way, a new dialogue, by which* P *defends the initial claim. To this end,* P *applies backward move on the current dialogue to find a new dialogue. The idea of the backward move is as follows.*

- *First,* P *tries to find a new forward move different from* v_2*. This attempt is failed because*

$\varphi_c^{v_1}$ is unsatisfiable and c is assigned to **t** *in the challenge move v_1. Then, P goes one step back and asks O to present a new challenge over the initial claim except the one which is in v_1.*

- *O checks the acceptance condition of b, $\varphi_b^{v_0} : \neg a \vee c$, and says that b can also be accepted if a is denied. Thus, the forward move is $v_1' =$* **ftu**.
- *Since $v_0 <_i v_1'$ the dialogue continues. Since v_1' is equal with the v_1 in the beginning of the example, by presenting v_2, P wins the dialogue and the game, as well.*

3 Discussion

Most argumentation frameworks are based on abstract argumentation, which determines an argument's acceptability. However, the role of discussion, which is a main feature of argumentation, has not been clarified in most of the abstract formalisms. As a part of my PhD, I clarify the role of discussion on semantics of ADFs by presenting discussion games.

The first existing game for answering the credulous problem of ADFs has been presented in [Keshavarzi Zafarghandi et al., 2019], my recent publication, which focuses on preferred semantics. I am working on the modification of that game in which the forward move is adjusted and a new backward move is defined to reduce the computational complexity of the game in the best case and in the average case. Both games answer the credulous decision problem of ADFs under preferred semantics. Further, both games work locally on the truth value of arguments which are claimed/challenged, that is, both try to find the truth values over parents of the arguments which are claimed/challenged.

In the new version of the game, the player whose turn it is uses the information which is presented by the competitor in the directly preceding claim/challenge move v, by computing φ_a^v for argument a, that is claimed/challenged in v. Then, the player looks for the truth values of arguments in φ_a^v to satisfy $v(a)$. However, in the first version, this step had to be done over all parents of a.

Note that since the acceptance conditions of arguments are presented by propositional formulas, it is possible that there exist more than one sets of truth values over parents of a that satisfy a claim/challenge. For instance, in Example 2.1, the acceptance condition b, namely $\varphi_b : \neg a \vee c$ says that b can be accepted in an interpretation if either $w_1 = \{a \mapsto \mathbf{f}\}$, $w_2 = \{c \mapsto \mathbf{t}\}$ or $w_3 = \{a \mapsto \mathbf{f}, c \mapsto \mathbf{t}\}$. In the current version, after picking a w_i over parents of b in $\varphi_b^{v_0}$, the player continues applying a forward move on v_0 and w_i, to revise the information of v_0. However, in the previous version, O first collected the set $W = \bigcup_{i=1}^{3}\{w_i\}$. In general, the number of elements of W can be blown up to 2^m, where m is the number of parents of a. Thus, the previous method has higher best case and average case computational complexity than the new version.

On the other hand, in the previous method, if dialogue $D = [v_0, \ldots, v_n]$ faces with a contradiction, if n is even, then P picks another element of W, for instance w' that P did not use before, and applies a forward move on v_{n-1} and w', and if n is odd O does the same. In the current version to overcome the lack of the set W, I defined a backward move which is applied by P on D to find a new dialogue.

I am working on an algorithm based on the game presented in Section 2. It appears that this algorithm can also be used as tool in human-machine interaction. As a future work, I will provide a solver based on this method and do an experiment to compare the performance of different solvers of ADFs in the credulous decision problem for preferred semantics.

4 Conclusion

In my current work, preferred discussion games between two agents, proponent and opponent, are considered as a proof method to investigate credulous acceptance (denial) of arguments in an ADF under preferred semantics. The presented methodology can be reused in AFs and generalizations of AFs that can be represented as subclasses of ADFs. Winning *one* dialogue of the game by P is sufficient to show that there exists a preferred interpretation in which the initial claim is satisfied. When there is a preferred

interpretation that satisfies the initial claim, via the current method, in the best case and even in the average case, there is no need to enumerate all preferred interpretations of an ADF to answer the credulous problem. The method is sound and complete.

Acknowledgements

Supported by the Center of Data Science & Systems Complexity (DSSC) Doctoral Programme, at the University of Groningen.

References

[Booth et al., 2018] Booth, R., Caminada, M., and Marshall, B. (2018). DISCO: A web-based implementation of discussion games for grounded and preferred semantics. In *Proceedings of Computational Models of Argument COMMA*, pages 453–454. IOS Press.

[Brewka and Woltran, 2010] Brewka, G. and Woltran, S. (2010). Abstract dialectical frameworks. In *Proceedings of the Twelfth International Conference on the Principles of Knowledge Representation and Reasoning (KR 2010)*, pages 102–111. AAAI Press.

[Caminada, 2017] Caminada, M. (2017). Argumentation semantics as formal discussion. *Handbook of Formal Argumentation*, 1:487–518.

[Dung, 1995] Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357.

[Dung and Thang, 2007] Dung, P. M. and Thang, P. M. (2007). A sound and complete dialectical proof procedure for sceptical preferred argumentation. In *Proc. of the LPNMR-Workshop on Argumentation and Nonmonotonic Reasoning (ArgNMR07)*, pages 49–63.

[Keshavarzi Zafarghandi et al., 2019] Keshavarzi Zafarghandi, A., Verbrugge, R., and Verheij, B. (2019). Discussion games for preferred semantics of abstract dialectical frameworks. In *European Conference on Symbolic and Quantitative Approaches with Uncertainty*, pages 62–73. Springer.

[Prakken and Sartor, 1997] Prakken, H. and Sartor, G. (1997). Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-classical Logics*, 7(1-2):25–75.

[Strass and Wallner, 2015] Strass, H. and Wallner, J. P. (2015). Analyzing the computational complexity of abstract dialectical frameworks via approximation fixpoint theory. *Artificial Intelligence*, 226:34–74.

Algorithms and Tools for Abstract Argumentation

Mickaël Lafages

IRIT, University of Toulouse, France

Abstract

Computing acceptability semantics of abstract argumentation frameworks is receiving increasing attention. Focused on finding algorithms and tools to make the abstract argumentation domain progress, this paper presents the objective of my PhD. So far, a distributed and clustering based algorithm, *AFDivider*, has been proposed. Designed for Dung's argumentation framework, it enumerates the acceptable sets of the main semantics proposed by Dung. Empirical results are presented. Possibility of extensions to other more expressive argumentation frameworks are planned.

1 Introduction

Among several approaches dealing with argumentation, Abstract Argumentation Theory proposes methods to represent and deal with contentious information, and to draw conclusions or to take decision from it. It is called "abstract" because it does not focus on how to construct arguments but rather on how arguments affect each other. Arguments are seen as generic entities that interact positively (support relation) or negatively (attack relation) with each other.

At first glance, such an approach may seem to be only theoretical but this abstraction level allows to propose generic reasoning processes that could be applied to any precise definition or formalism for arguments. Argumentation-based reasoning model has been of application in multi-agent systems for years now (see [Carrera and Iglesias, 2015] for an overview). The development of argumentation techniques and of their computation drives such applications. This is the very motivation of my PhD studies: enhancing the use of abstract argumentation, and more generally argumentation, by developing better tools, especially algorithms.

A lot of "frameworks" have been designed to enhance expressivity in abstract argumentation (*e.g.* [Nouioua and Risch, 2010, Baroni et al., 2011, Coste-Marquis et al., 2012, Amgoud et al., 2017]) as well as "semantics". While a given framework specifies the way of representing and expressing an argumentation problem (types of relations between arguments, weight on attacks or arguments, higher-order relation, etc.), a semantics, defined for a specific argumentation framework (AF), captures what is a solution of an argumentation problem, in the sense of what is acceptable.

The study roadmap of my PhD is to first focus on solving more efficiently argumentation problems that are expressed in the basic, seminal argumentation framework and semantics defined by Dung [Dung, 1995]. Then, the idea is to extend my work for more enriched argumentation frameworks.

Dung's semantics produce sets of arguments, so-called "extensions". Those arguments, taken together, are solution of the argumentation problem. The main contribution of my PhD so far is the proposal of a new distributed and clustering based algorithm to compute Dung's semantics. It has been designed for certain types of "large-scale" argumentation frameworks, that produce a lot of different extensions.

The principles of this algorithm are discussed

in Section 2. In Section 3 the results of this application are shown and possible extensions to other frameworks are presented. Perspective for future work is then opened.

2 A distributed and clustering based algorithm

The idea that leads to the so-called *AFDivider* algorithm is that one could take advantage of the shape of the argumentation framework to compute more efficiently the extensions of a given semantics. Let consider argumentation frameworks with space and dense areas. Rather than building extensions that cover the whole AF, which could be time consuming, it may be a good idea to cut the AF into pieces along those identified dense areas, compute simultaneously parts of extensions and finally wisely reunify extensions parts together. This is the big picture of the proposed algorithm. We are going to see in more details how it works.

The first step of the *AFDivider* algorithm is to remove the so-called "trival part" of an AF. Given that the *grounded* extension is included in all *complete*, and so *preferred* and *stable* extensions and that the *grounded* extension can be found in linear time, we first compute it and remove all the arguments concerned by it from the AF (all arguments that are in it or attacked by it). The resulted AF is considered as the "hard part" of the AF. Notice that it may be non connected. The *AFDivider* algorithm takes advantage of it.

The second step is to find clusters in the resulting AF. There exist several algorithms to cluster graphs. The *AFDivider* algorithm uses a spectral clustering method, usually used in machine learning for cluster identification. It is particularly well suited for sparse graph and this fits the type of AF we are interested in. Without going into details, this clustering is based on a similarity measure between arguments. In our case it is the number of relations between neighbouring arguments. From the overall pairs of similarity, a matrix computation is done in order to find well shaped clusters as much as possible.

The third step is to compute parts of extensions according to a given semantics (*complete*, *preferred* or *stable*). For each argument in a given cluster which is attacked by an argument outside this cluster, we compute the semantics of this cluster by considering that its attackers could be accepted, rejected or in an undecidable state in their own clusters. This computation is made in a simultaneous way.

The last step is to reunify the parts of extensions together. The cluster parts are reunified with respect to the constraints on cluster external relation states. Then parts of connected components are joined together (this does not require constraint checking as there are no relation between connected components). Finally the *grounded* part is added to all of them. This is how we obtain extensions of the whole AF. [1]

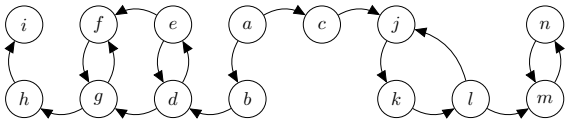
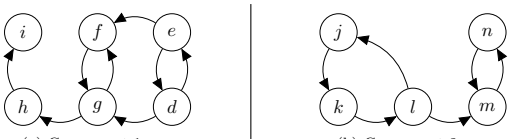


Figure 1: Example of an argumentation framework: Γ

Let briefly illustrate the *AFDivider* algorithm on the AF shown in Figure 1 for the *complete* semantics.

Step 1: The grounded extension is $\{a\}$: a, b and c are removed from Γ with the attacks involving them. We obtain two connected components as show in Figure 2.



(a) Component 1: γ_1. (b) Component 2: γ_2.

Figure 2: Connected components resulting from the grounded removal pre-processing.

Step 2: Four clusters are determined from γ_1 and γ_2: κ_1, κ_2, κ_3 and κ_4 as shown in Figure 3.

Step 3: The cluster extensions are computed simultaneously. We have:

[1] Note that some subtleties for the computation of the *stable* and *preferred* semantics are not presented for sake of brevity.

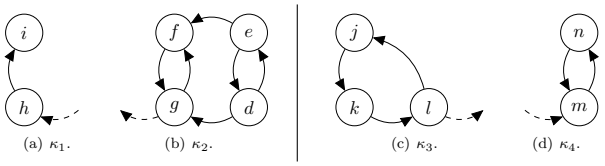


Figure 3: Identified clusters.

- For κ_1, three extensions: $\{h\}$, $\{i\}$ and $\{\}$.
- For κ_2, three extensions: $\{d, f\}$,$\{e, g\}$ and $\{\}$.
- For κ_3, one extension: $\{\}$.
- For κ_4, three extensions: $\{m\}$, $\{n\}$ and $\{\}$.

Step 4: Finally, the reunifying process checking compatible parts produces six extensions: $\{a\}$, $\{a, n\}$, $\{a, d, f, h\}$, $\{a, d, f, h, n\}$, $\{a, e, g, i\}$ and $\{a, e, g, i, n\}$.

3 Discussion

A competition, ICCMA, that compares argumentation solvers on their ability to solve the enumeration of extensions problem (and other decision problems) was created a few years ago.[2] Some editions of this competition have been analyzed: [Bistarelli et al., 2018, Rodrigues et al., 2018] highlight that some AF instances have been particularly hard to solve, and that others were not solved at all, considering the *preferred* semantics notably. Many of these instances are of Barabási–Albert (BA) type [Albert and Barabási, 2002], which is a structure found in several large-scale natural and human-made systems, such as the World Wide Web and some social networks [Barabási et al., 2016]. Those types of AF are among the ones *AFDivider* has been designed for.

In order to evaluate the performances of our algorithm we compare it with some of the best solvers[3] at that time for the *complete*, *stable* and the *preferred* semantics and on some hard AF instances that have been identified.

For each experiment, we used a 6 core processor, each core having a frequency of 3 GHz. The RAM size was 45GB. The timeout had been set to 1 hour.

The results, reported in Table 1, show that this approach of clustering and reunifying extension parts is very relevant for some types of AF.[4] Indeed, although for the *stable* semantics *pyglaf* and *AFDivider* have similar solving time, for the *preferred* semantics we can observe a real change of order of magnitude.

These experiments led to publications. See [Lafages et al., 2018] and [Doutre et al., 2019] for deeper explanations and analysis[5].

Further analysis for other clustering methods and with new solvers are currently in progress. A work to propose algorithms for other frameworks is also in process. The first framework types we are interested in are the ones with higher-order attacks, that is, that allow attack on attacks, especially AFRA (see [Baroni et al., 2011]) and RAF (see [Cayrol et al., 2017]). Before proposing algorithms, a work must be done for determining the complexity of computing semantics in those frameworks. There is also a need of tools such as labellings[6] that is more convenient for an algorithmic approach than searching for sets of elements. A preliminary part of these works can already be seen, as published technical reports: [Doutre et al., 2020b]

[2] International Competition on Computational Models of Argumentation (ICCMA) `http://argumentationcompetition.org/`.

[3] See respectively [Alviano, 2018] and [Cerutti et al., 2017] for details on *Pyglaf* and *ArgSemSAT* solvers.

[4] *amador-transit_20151216_1706.gml.80.apx* and *basin-or-us.gml.20.apx* are instances which come from real data of the traffic domain. In Table 1, i_1 to i_8 correspond respectively to BA_120_70_1.apx, BA_100_60_2.apx, BA_120_80_2.apx, BA_180_60_4.apx, basin-or-us.gml.20.apx, BA_100_80_3.apx, amador-transit_20151216_1706.gml.80.apx and BA_-200_70_4.apx. Note that these instances have a number of extensions under the *preferred* and *stable* semantics that is particularly large (more than a hundred thousand), and even larger for the complete semantics.

[5] Note that in Table 1, `FAIL` means that the given solver failed to solve the problem due to the limit of time or to the limit memory space. For more details see the mentioned publications.

[6] Labelling is a three value mapping which associates to each argument of an AF a status, accepted, rejected or undecidable.

		Instances							
		i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
PR	Nb ext. ($\approx$)	0.28×10^6	1.07×10^6	1.28×10^6	1.37×10^6	1.96×10^6	4.47×10^6	11.75×10^6	10.74×10^9
	AFDivider	0:05.84	0:27.98	0:20.42	0:35.05	0:31.31	1:09.10	12:39.21	FAIL
	Pyglaf	0:39.00	6:04.37	10:12.22	14:51.09	54:20.72	FAIL	FAIL	FAIL
	ArgSemSAT	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL
ST	Nb ext. ($\approx$)	Idem preferred case							
	AFDivider	0:06.26	0:13.20	0:18.78	0:31.02	0:29.46	0:50.79	1:48.30	FAIL
	Pyglaf	0:03.02	0:09.22	0:14.76	0:18.43	0:21.15	0:42.57	1:53.95	FAIL
	ArgSemSAT	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL
CO	Nb ext. ($\approx$)	0.80×10^9	5.22×10^9	9.31×10^9	11.93×10^9	16.18×10^9	49.58×10^9	-	22×10^{15}
	All three solvers	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL	FAIL

Table 1: Experimental results (PR: preferred, CO: complete, ST: stable,"-": "missing data"). The time result format is "minutes:seconds.centiseconds".

and [Doutre et al., 2020a].

4 Conclusion

As a conclusion, my PhD studies focus on finding algorithms and tools to make the abstract argumentation domain progress. A very interesting approach to compute semantics has been proposed so far and works are in progress to extend this solving method, or other ones, to more expressive argumentation frameworks.

Acknowledgements

I thank my PhD supervisors Marie-Christine Lagasquie-Schiex and Sylvie Doutre for their genuine advices and their support.

References

[Albert and Barabási, 2002] Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

[Alviano, 2018] Alviano, M. (2018). The pyglaf argumentation reasoner. In *OASIcs-OpenAccess Series in Informatics*, volume 58. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[Amgoud et al., 2017] Amgoud, L., Ben-Naim, J., Doder, D., and Vesic, S. (2017). Acceptability semantics for weighted argumentation frameworks. In *Proceedings of IJCAI*, volume 2017.

[Barabási et al., 2016] Barabási, A.-L. et al. (2016). *Network science*. Cambridge university press.

[Baroni et al., 2011] Baroni, P., Cerutti, F., Giacomin, M., and Guida, G. (2011). Afra: Argumentation framework with recursive attacks. *International Journal of Approximate Reasoning*, 52(1):19–37.

[Bistarelli et al., 2018] Bistarelli, S., Rossi, F., and Santini, F. (2018). Not only size, but also shape counts: abstract argumentation solvers are benchmark-sensitive. *J. Log. Comput.*, 28(1):85–117.

[Carrera and Iglesias, 2015] Carrera, Á. and Iglesias, C. A. (2015). A systematic review of argumentation techniques for multi-agent systems research. *Artificial Intelligence Review*, 44(4):509–535.

[Cayrol et al., 2017] Cayrol, C., Fandinno, J., Fariñas del Cerro, L., and Lagasquie-Schiex, M.-C. (2017). Valid attacks in argumentation frameworks with recursive attacks. In *13th International Symposium on Commonsense Reasoning (Commonsense)*, volume 2052. CEUR-WS : Workshop proceedings.

[Cerutti et al., 2017] Cerutti, F., Vallati, M., Giacomin, M., and Zanetti, T. (2017). ArgSemSAT-2017.

[Coste-Marquis et al., 2012] Coste-Marquis, S., Konieczny, S., Marquis, P., and Ouali, M. A. (2012). Weighted attacks in argumentation frameworks. In *KR*.

[Doutre et al., 2019] Doutre, S., Lafages, M., and Lagasquié-Schiex, M.-C. (2019). A distributed and clustering-based algorithm for the enumeration problem in abstract argumentation. In *International Conference on Principles and Practice of Multi-Agent Systems*, pages 87–105. Springer.

[Doutre et al., 2020a] Doutre, S., Lafages, M., and Lagasquie-Schiex, M.-C. (2020a). Argumentation Frameworks with Higher-Order Attacks: Complexity results. Rapport de recherche IRIT/RR–2020–03–FR, IRIT, Université Paul Sabatier, Toulouse.

[Doutre et al., 2020b] Doutre, S., Lafages, M., and Lagasquie-Schiex, M.-C. (2020b). Argumentation Frameworks with Higher-Order Attacks: Labelling Semantics. Rapport de recherche IRIT/RR–2020–01–FR, IRIT, Université Paul Sabatier, Toulouse.

[Dung, 1995] Dung, P. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.

[Lafages et al., 2018] Lafages, M., Doutre, S., and Lagasquie-Schiex, M.-C. (2018). Clustering and distributed computing in abstract argumentation. Rapport de recherche IRIT/RR–2018–11–FR, IRIT, Université Paul Sabatier, Toulouse.

[Nouioua and Risch, 2010] Nouioua, F. and Risch, V. (2010). Bipolar argumentation frameworks with specialized supports. In *Tools with Artificial Intelligence (ICTAI), 2010 22nd IEEE International Conference on*, volume 1, pages 215–218. IEEE.

[Rodrigues et al., 2018] Rodrigues, O., Black, E., Luck, M., and Murphy, J. (2018). On structural properties of argumentation frameworks: Lessons from ICCMA. In *SAFA workshop*, pages 22–35.

Crafting neural argumentation networks

Jack Mumford

Department of Informatics, King's College London, UK

Abstract

This paper presents work on constructing neural network architectures, defined as neural argumentation networks (NANs), such that learning is conducted according to argumentation principles. NANs are designed to learn attack-relations that are consistent with input argument acceptability data. Hence this work describes a process of calculating legitimate attack-relations that does not rely on examining the structure of the arguments, which are not always clear. The paper describes the theory and application of translating argumentation semantics, as outlined by Dung, to NAN architectures and compares two distinct forms: 3-valued acceptability, and 2-valued acceptability. Future work can be envisioned as extending the methodology to incorporate more complex argumentation semantics including support and weighted approaches that could align with wider data-sets and reasoning problems.

1 Introduction

Neural argumentation networks (NANs) are computational neural networks that learn according to argumentation principles. Argumentation, as a field within artificial intelligence, is highly relevant for researchers interested in symbolic representations of knowledge and defeasible reasoning [Bench-Capon and Dunne, 2007]. Argumentation is proffered as a semantics-based logic that models human reasoning whilst retaining mathematical integrity. Argumentation frameworks (AFs) [Dung, 1995] offer a graph-based approach that determines logically consistent argument positions solely via attack-relations between the arguments. The shared graphical nature of AFs and computational neural networks make their union a sensible choice when looking to bridge the potentially explainability-intractable numerical methods with high level symbolic interpretation through concepts.

We show that there exists a one-one mapping between NANs and AFs which is determined by the attack-relations between arguments. Thus when a NAN learns, it learns attack-relations according to input argument acceptability data. Argument acceptability data comes in the form of a set of labellings, with each argument's label value indicating its acceptability within a particular labelling. Each labelling represents a single consistent position, such as an individual's opinion on the acceptability of the arguments. There is a set of labellings when there is a data set of many positions, e.g. in a debate between many individuals. But NANs are not limited to reasoning with human opinions on arguments, as any data for which each data point could be interpreted as an argument with an acceptability label would be relevant.

If an AF is already known in full, with fixed attack-relations, then no learning is required and the NAN would be used solely for calculating argument acceptability from input data. However, attack-relations will not always be known *a priori* and so would need to be learned. The research in this paper focuses explicitly on the learning aspect of the NAN architecture. This can be described as the reverse of the traditional argumentation problem, which deduces argument acceptability according to input attack-relations. Learning attack-relations from argument

acceptability data is especially relevant when the structure of the arguments does not reliably reveal the true attacks. An example of this is when an enthymeme's unstated premise is attacked; the structure is hidden and so the attack is also hidden.

The next three parts of this paper detail the novel contributions made in achieving the learning of attack-relations from input argument acceptability data:

1. Learning attack-relations from argument acceptability data in the form of 3-valued argument labellings.
2. Applying gradualism to the learning process and accommodating noisy data.
3. Learning attack-relations from argument acceptability data in the form of 2-valued argument labellings.

The ordering above indicates the sequence of reasoning, in that each subsequent part builds upon work conducted previously. The third section offers a discussion of the results and complexity of the various learning algorithms. A concluding section will then summarise the current progress and offer suggestions for future expansions.

2 3-valued attack learning problem

Reinstatement labellings [Caminada, 2006] form the basis for the 3-valued labelling approach, in which arguments will be labelled IN, UNDEC or OUT in any given labelling. In addition, a special label of *null* is used for those arguments that are not labelled in the particular labelling but are labelled for some other labelling. In other work awaiting publication, it is shown that the attack learning problem (in which acceptability data is provided as input and attacks are produced as output) is significantly less complex for 3-valued labellings than for 2-valued labelling if the objective is to find an attack set that is fully consistent with the input data. It is revealed that although both labelling types are solvable within polynomial-space complexity, the time complexity is very different. Polynomial-time algorithms exist for the 3-valued labelling approach whereas the 2-valued labelling approach is proven to be NP-complete. Moreover, three labels allow for less ambiguous reasoning by providing greater distinction between argument acceptability statuses. The increase in reasoning precision for 3-valued labels would intuitively produce more accurate attack-relations that fit the data less ambiguously than the 2-valued variety. Combination of the two incentives of reduced complexity and conjectured higher accuracy, forms the basis for adopting the 3-valued attack learning problem as first choice for NAN implementation.

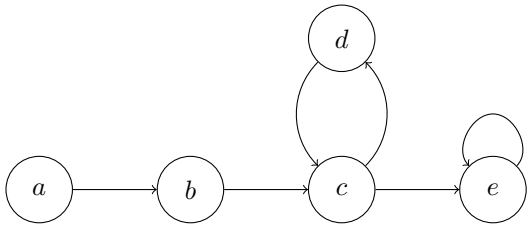


Figure 1: A framework containing a non-empty grounded extension $\{a\}$, two preferred extensions $\{(a, c), (a, d)\}$, one stable extension $\{a, c\}$ an even cycle, and two odd cycles including one self-attacking argument

Each NAN architecture forms a one-one mapping with some AF. For example, the framework in Fig. 1 is converted into the NAN depicted in Fig. 2. Only a simple one-layer neural network is required, but with essential added constraints on the learning process that account for incongruity that would result from the application of backpropagation alone. Several algorithms were developed for the purpose, which are detailed in other work pending publication. Although the algorithms differ in function, they all operate by forward processing each labelling within the dataset, producing output argument labels based on the labels of their attacking arguments according to complete labelling semantics. The backward learning process then compares the output argument labels to the target labels, which are the same as the input labels, and adjusts the weighted edges, and hence the attacks, according to the error.

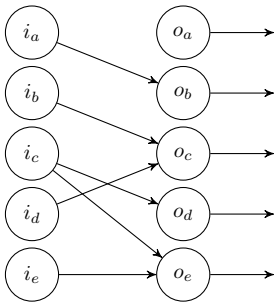


Figure 2: A neural network processing 3-valued labellings, corresponding to the framework detailed in Figure 1. An attack (a, b) exists *iff* an arrow joins node i_a to node o_b.

3 Gradualism and noisy data algorithms

The previous section detailed the process of learning attacks from 3-valued argument acceptability data. However, the learning progressed such that networks edges, representing attacks, were not incrementally updated but rather adjusted in a binary manner such that an attack was true or false according to the learning error.

By incorporating gradualism into the NAN architecture, the shape of a given data set has more affect on the attack-relation learned. Thus, by setting an expedient learning rate, attacks are adjusted in proportion to the amount of error corrected. This approach helps to create sparser attack-relations, which are more appropriate for domains in which it is assumed that attacks between arguments are rare. Again, several algorithms have been developed that employ gradualism to address the 3-valued attack learning problem.

The final work on the 3-valued approach extends the incremental learning of the gradualism algorithms to accord with noisy data. The previous algorithms all operated by searching for an exact solution equivalent to an AF that would be completely consistent with the input argument acceptability data set in keeping with complete labelling semantics. The resulting algorithms are more aligned with general machine learning in which zero error is not assumed to be possible. Instead of finding a zero error solution, the learning is focused on minimising the error.

4 2-valued attack learning problem

Directly translating complete extension semantics to a labelling form gives rise to two labels: IN and NOTIN, denoting an argument is within a particular extension or not respectively. Again the special label of *null* is required for arguments without a label for a given labelling. The 2-valued attack learning problem addresses the same scenario as for the 3-valued attack learning problem, but with the 2-valued form of labelling as input argument acceptability data.

Interestingly, the shift in focus necessitated when handling noisy data is also employed for the 2-valued problem. The NP-complete complexity required for an absolute solution attack-relation is clearly computationally expensive. Hence, by redefining the problem to minimising the labelling errors, bounded by a stopping criteria, it is possible to construct a NAN architecture and associated algorithms that are more tractable.

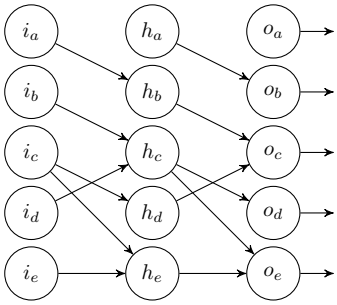


Figure 3: A neural network processing 2-valued labellings, corresponding to the framework detailed in Figure 1. An attack (a, b) exists *iff* an arrow joins node i_a to node h_b (which also dictates that an arrow joins node h_a to node o_b).

Fig. 3 depicts the basic structure of the 2-valued NAN architecture. An additional layer is required in order to effectively process the ambiguity presented by the NOTIN label which can correspond to either UNDEC or OUT in the 3-valued labelling approach.

The hidden layer is used to convert the 2-valued input to the 3-valued approach so that the output is consistent with the fix-point requirements of complete semantics. The learning is constrained so that for any given argument a and an argument b the value of the edge connecting i_a to h_b is the same as that connecting h_a to o_b. This ensures there is no ambiguity with respect to which edge should denote the attack relation, but requires learning updates to apply to the two edges uniformly.

5 Discussion

As referenced earlier, the complexities of solving the attack learning problems vary depending on what type of data labelling is provided as input, the algorithm concerned, and whether a zero-error solution is required. Table 1 shows that if the objective is to find an attack set fully consistent with input data then the time complexity of the 2-valued problem is significantly worse than for the 3-valued problem. It is interesting to note that these results also hold for algorithms implementing gradualism. However, when noise is introduced, the problem objective switches to optimization and it is conjectured that a solution for even the 3-valued problem would be NP-complete. This is why we require alternative stopping criteria that allow for a 'good-enough' solution in lieu of a prohibitively expensive actual solution.

Problem	**Time complexity**	**Space complexity**
3-valued	$O(n^2\|T\|)$	$O(n\|T\| + n^2)$
2-valued	NP-complete	$O(n\|T\| + n^2)$

Table 1: Best-case complexities for the 3-valued and 2-valued attack learning problems for finding zero-error solution.

Beyond theory, an empirical study of all algorithms was performed with the use of data from the ICCMA2017 competition, which provided examples of AFs and associated argument extensions. The algorithms were trained on the argument labellings pertaining to the extension sets and the output attack-relations were compared to the 'actual' attack-relations from the original data to measure the fidelity of the results. In summary, high accuracy results were obtainable across all algorithms for the overall attack-relation (a binary classification in which all possible attacks were classed in or out), but attack precision proved to be the greatest challenge. That said, the performance of 3-valued algorithms was superior in all metrics compared with 2-valued alternatives. Algorithms favoring sparse attack-sets also performed better, especially in attack precision, which reflects the rarity of attacks in the actual ICCMA2017 data.

Both the complexity results and the empirical study are proposed as novel contributions to the field. Similar work on realizability [Dunne et al., 2015] and argumentation synthesis [Niskanen et al., 2019], address the attack learning problem but do not succeed in achieving complexity results for complete semantics. In addition, both related works examine attack learning through different lens; realizability requires the solution attack-relation to map one-one to the input acceptability data, argumentation synthesis is focused on minimising the number of errors (akin to the noisy data algorithms referred to in this paper) but frames the problem as a traditional MAX-SAT search problem.

Future work can be anticipated in manifold directions. Expanding the empirical evaluation to real applications beyond the more abstract data obtained from ICCMA2017 is vital to assess the potential domains for which NANs can have significant impact in identifying reliable attack-relations. From a theoretical viewpoint, NAN algorithms could be developed for more advanced semantics that, for example, incorporate weighted approaches and support into their mechanics, and empirical evaluation could be performed on such algorithms.

6 Conclusion

This paper has described the concept of neural argumentation networks (NANs) that address the attack learning problem of calculating an attack-relation based on consistency with input argument acceptability data. Three broad novel contributions were identified with respect to:

1. Learning attack-relations from argument acceptability data in the form of 3-valued argument labellings.
2. Applying gradualism to the learning process and accommodating noisy data.
3. Learning attack-relations from argument acceptability data in the form of 2-valued argument labellings.

The advantages of reduced complexity and superior attack-relation fidelity derived from using 3-valued data, as opposed to 2-valued data, were evinced from both theoretical and empirical results.

The research is relevant in the scope that attack-relations are essential for the construction of argumentation frameworks. The NAN algorithms are proffered as a means of calculating such attack-relations when prior knowledge and extraction from the argument structure is not reliable. Further research is needed to examine the impact of the algorithms with real domains and to evaluate the effects of enrich the algorithms with alternative mechanisms such as support and weighted argumentation.

Acknowledgements

The research described by this short paper could not have been possible without their influence and/or contribution: E. Black, K. Mumford, S. Parsons and I. Sassoon.

References

[Bench-Capon and Dunne, 2007] Bench-Capon, T. J. and Dunne, P. E. (2007). Argumentation in artificial intelligence. *Artificial intelligence*, 171(10-15):619–641.

[Caminada, 2006] Caminada, M. (2006). On the issue of reinstatement in argumentation. In *European Workshop on Logics in Artificial Intelligence*, pages 111–123. Springer.

[Dung, 1995] Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.

[Dunne et al., 2015] Dunne, P. E., Dvořák, W., Linsbichler, T., and Woltran, S. (2015). Characteristics of multiple viewpoints in abstract argumentation. *Artificial Intelligence*, 228:153–178.

[Niskanen et al., 2019] Niskanen, A., Wallner, J., and Järvisalo, M. (2019). Synthesizing argumentation frameworks from examples. *Journal of Artificial Intelligence Research*, 66:503–554.

Understanding Stories Using Crowdsourced Commonsense Knowledge

Christos T. Rodosthenous

Open University of Cyprus

Abstract

This paper presents work on automated story understanding by using commonsense knowledge acquired from human contributors. A description of the methodology followed for acquiring this knowledge is depicted, followed by a presentation on how argumentation is used for representing commonsense knowledge. Furthermore this work includes a presentation of tools that are designed and developed to acquire and apply knowledge for the purpose of understanding stories.

1 Introduction

One of the major problems in Artificial Intelligence, that is still an obstacle in the development of "intelligent" machines [McCarthy, 1959], is the lack of commonsense knowledge. This work focuses on the problem of commonsense knowledge acquisition and the application of this knowledge on the task of story understanding. Humans are able to understand a text passage quite easily, starting from young ages, but machines face a lot of difficulties for the same task. One of the prerequisites for text understanding is the existence of commonsense knowledge, i.e., knowledge that is not explicitly present but it is inferred. Examples of such knowledge can be easily extracted from our daily lives, such as "when the sun is up, its daytime","if you just woke up then you should have been sleeping" and many more.

Why stories? This work focuses on stories, since they are a very good example of texts with structure, plot, events that change over time and they hold a predominant position in the learning path of humans. From the early childhood, most parents read fairy tales to their kids both for fun and for learning. The importance of story understanding is highlighted by Winston, who stated that "story understanding is the centrally important foundation for all human thinking" [Winston, 2012] .

There were various attempts to gather commonsense knowledge either by using experts in logic (e.g., CyC [Lenat, 2019]) or by using the crowd (e.g., the Open Mind Commonsense Project [Singh et al., 2002]). The latter gained a lot of attention with the introduction of the internet, where access to the crowd was made much easier than before. Lately, we have seen an increasing interest from individual researchers and institutions to gather commonsense knowledge.

The approach taken in this work, is that commonsense knowledge appropriate for story understanding can be gathered by sourcing the task to humans, using crowdsourcing [Howe, 2006], where both intrinsic and extrinsic methods for knowledge acquisition are employed.

Knowledge acquisition is only one part of the problem that this work needs to address. There is also the need to find an appropriate representation for the acquired knowledge. There are attempts to represent knowledge in a variety of ways, using strict logic based rules, scripts, and graphs. Singh et al. (2002) suggested that this knowledge should be represented in natural language but this adds another obstacle in the utilization of the acquired knowl-

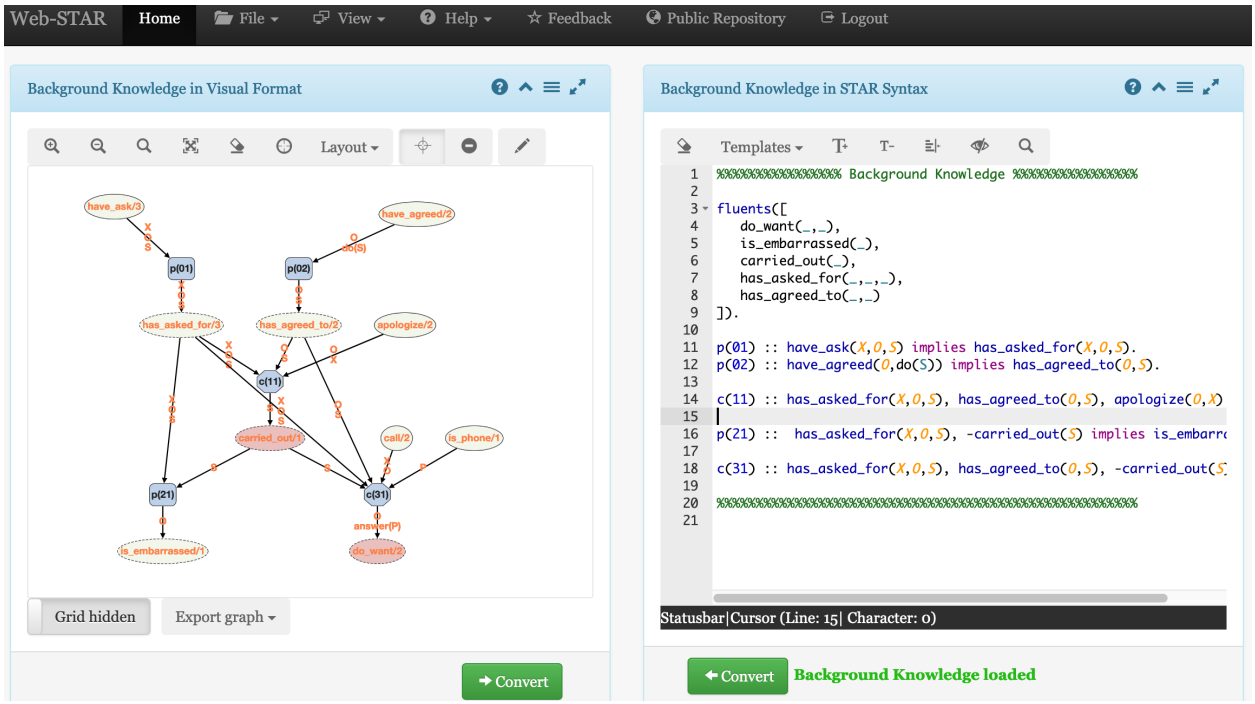


Figure 1: A screenshot of the Web-STAR interface depicting background knowledge representation, both visually (left side) and textually (right side).

edge from machines. This is the part where computational argumentation kicks in. Argumentation is an appropriate substrate for representing knowledge since it fits the incomplete and conflicting nature of commonsense knowledge [Kakas and Michael, 2016, Michael, 2019]. Moreover, argumentation is also suitable to handle preferences between commonsense knowledge. Argumentation semantics are used both for knowledge representation and for reasoning in the internal mechanisms of the tools developed for story understanding, as part of this work.

2 Method

Research work starts with a literature review on the current state of affairs on story understanding, commonsense knowledge acquisition and appropriate representations of the acquired knowledge. This step provides insights on what was already achieved in these fields and outlines possible obstacles identified by other researchers. During this stage, a number of systems were identified but the focus of the work is on the ones that use human computation or crowdsourcing as a method for acquiring knowledge.

The chosen methodology centers on testing both implicit and explicit knowledge acquisition methods. *Implicit acquisition methods* are described as the ones that knowledge is contributed without contributors actively knowing that they are performing this task, but it is actually a hidden/side task. *Explicit acquisition methods* are described as the ones in which contributors know that they are contributing knowledge and it is their primary task. In this category the motives are either monetary or research/learning/social oriented. For the former, motives are mostly fun since these methods most of the time include the use of games, i.e., Games With A Purpose (GWAPS) [von Ahn and Dabbish, 2008].

Firstly, a tool for facilitating users to encode a story, and to manually add knowledge rules in a way that machines can understand them is presented. This tool is a Web-based Integrated Development Environment (IDE) called "Web-STAR" [Rodosthenous and Michael, 2019b]. It facilitates both expert and non-expert users in encoding stories in symbolic form and adding background knowledge (cf. Figure 1), providing also a number of embedded utilities for converting natural language stories to symbolic format, visually adding knowledge using a directed graph editor and collaboration functionality.

The internal mechanism of Web-STAR is based on the STory comprehension through ARgumentation (STAR) system [Diakidoy et al., 2015]. This is a prolog based system able to read a text file with story events at specific timepoints, a list of background knowledge rules in symbolic format, and a list of multiple-choice questions. The system outputs the comprehension model, i.e., which story event holds at each timepoint and it also provides answers to the questions posed. The STAR system uses a structured rule-based argumentation framework similar to that of ASPIC+ [Modgil and Prakken, 2014]. Combinations of premises from the story with defeasible rules from the background knowledge form a proof tree in support of some inference; this tree corresponds to an argument.

The output is presented both textually and graphically in a timeline format, where users can follow the comprehension model and track changes to the story timeline. The IDE was evaluated for its ease of use by both expert and non-expert users, following user experience measurement methodologies and it received a high score in its evaluation.

Secondly, a novel framework and platform [Rodosthenous and Michael, 2018a], developed for implementing crowdsourcing applications (e.g., Games with a Purpose or language learning applications) that can be used by human workers for gathering commonsense knowledge, is presented. Two experiments were designed and conducted, that examine whether fully automated or hybrid crowdsourcing techniques, i.e., techniques that benefit from both manually, crowd-contributed and automatic acquisition of knowledge, can be used to gather commonsense knowledge.

Two Games With A Purpose were developed using the aforementioned platform. The first is called "Knowledge Coder[1]" [Rodosthenous and Michael, 2014] and relies only on crowdsourcing approaches to acquire knowledge (cf. Figure 2). The second is called "Robot Trainer[2]" [Rodosthenous and Michael, 2016] and uses a hybrid methodology for gathering background knowledge (cf. Figure 3). More specifically, players generalize knowledge and evaluate its appropriateness in answering questions on unseen stories. The acquired knowledge was tested on story comprehension tasks, such as question answering. The results show that the gathered knowledge is useful in answering story questions on new unseen stories, since the gathered knowledge is applicable in stories other than the ones used to generate the knowledge.

Both GWAPs follow the approach of breaking this task into a sequence of more specific tasks, so that human participants not only identify relevant knowledge, but also convert it into a machine-readable form and evaluate its appropriateness.

Thirdly, the study sets out to investigate the problem of inferring the geographic focus of a story at a country level, i.e., the geographic location that the story is related to. An application was developed for inferring the geographic focus of news stories using crowdsourced knowledge bases, contributing in understanding the "Where" a story takes place type of question. This application, called "GeoMantis" [Rodosthenous and Michael, 2019a, Rodosthenous and Michael, 2018b] retrieves knowledge from popular crowdsourced knowledge bases, such as ConceptNet [Speer et al., 2017] and YAGO [Mahdisoltani et al., 2015] and returns a prediction of the country of focus.

Current research work is focused on expanding GeoMantis, by applying a crowdsourced strategy for this task where paid crowd-workers evaluate the usefulness of the arguments on identifying the geographic focus of a document and the evaluated arguments are tested on whether they improve the accuracy of iden-

[1] `https://cognition.ouc.ac.cy/knowledge_coder`
[2] `https://cognition.ouc.ac.cy/robot`

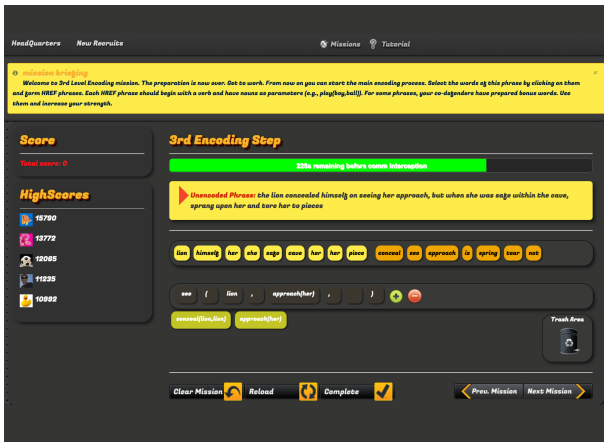


Figure 2: A screenshot of the 3rd mission from the **Knowledge Coder** game where players encode knowledge in logic form using verbs and nouns from a sentence.

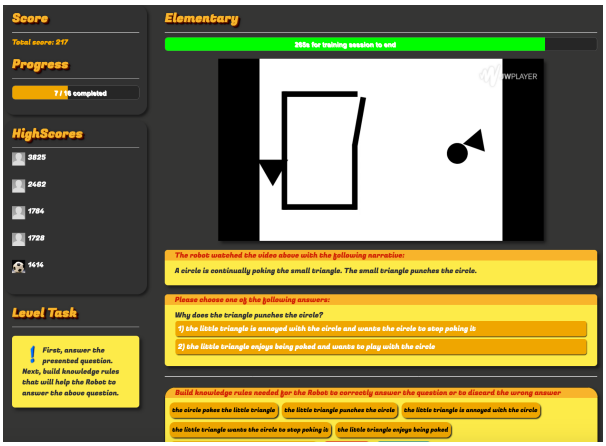


Figure 3: A screenshot of the 1st level from the **Robot Trainer** game where players generate rules by dragging and dropping short phrases in natural language to form appropriate knowledge rules.

tifying the geographic focus of stories. Results show encouraging indications that this hybrid methodology improves the accuracy.

3 Discussion and Conclusion

The aim of the present research was to investigate methods for acquiring and using commonsense knowledge for automated story understanding. A combination of crowdsourcing methods and computational argumentation allowed both the acquisition of relevant knowledge and its representation in a way that can be used for automated story comprehension. Computational argumentation, provides flexibility in handling knowledge, drawing inferences in the temporal setting of a story and based on evidences from psychology, it is compatible with how us humans comprehend a story [Diakidoy et al., 2014].

The usage of both implicit and explicit acquisition methods, and the use of crowdsourcing seems promising both in terms of quantity and quality of the acquired knowledge. Explicit acquisition methods require more time from contributors and expertise in encoding knowledge, whereas implicit ones can be used by more contributors and for longer period, but require more effort in designing the platforms/applications/games to facilitate the acquisition of knowledge.

In terms of applicability of crowdsourced knowledge for story understanding tasks, the acquired knowledge is suitable for question answering as it can be applied both to answer multiple-choice questions on unseen stories and to identify the geographic focus of news stories.

The tools presented in Section 2 are developed and have been tested in a number of scenarios. Moreover, tools like Web-STAR provide a friendly user interface for utilizing the capabilities of the underlying story understanding engine. Interested readers are directed to the relevant papers to get more information on each tool and how it is used.

Acknowledgements

This thesis is supervised by Dr Loizos Michael, Associate Professor at the Open University of Cyprus.

References

[Diakidoy et al., 2014] Diakidoy, I.-A., Kakas, A., Michael, L., and Miller, R. (2014). Story Comprehension Through Argumentation. In *Proceedings of the 5th International Conference on Computational Models of Argument (COMMA 2014)*, pages 31–42, Scottish Highlands, UK. IOS Press.

[Diakidoy et al., 2015] Diakidoy, I.-A., Kakas, A., Michael, L., and Miller, R. (2015). STAR: A System of Argumentation for Story Comprehension and Beyond. In *Working Notes of the 12th International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense 2015)*, pages 64–70.

[Howe, 2006] Howe, J. (2006). Crowdsourcing: A Definition.

[Kakas and Michael, 2016] Kakas, A. and Michael, L. (2016). Cognitive Systems: Argument and Cognition. *IEEE Intelligent Informatics Bulletin*, 17(1):14–20.

[Lenat, 2019] Lenat, D. (2019). Cyc Technology Overview White Paper.

[Mahdisoltani et al., 2015] Mahdisoltani, F., Biega, J., and Suchanek, F. M. (2015). YAGO3: A Knowledge Base from Multilingual Wikipedias. *Proceedings of CIDR*, pages 1–11.

[McCarthy, 1959] McCarthy, J. (1959). Programs With Common Sense. In *Proceedings of the Symposium on the Mechanization of Thought Processes*, pages 75–91. London.

[Michael, 2019] Michael, L. (2019). Machine Coaching. In *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI)*, pages 80–86.

[Modgil and Prakken, 2014] Modgil, S. and Prakken, H. (2014). The ASPIC+ Framework for Structured Argumentation: A Tutorial. *Argument & Computation*, 5(1):31–62.

[Rodosthenous and Michael, 2016] Rodosthenous, C. and Michael, L. (2016). A Hybrid Approach to Commonsense Knowledge Acquisition. In *Proceedings of the 8th European Starting AI Researcher Symposium (STAIRS 2016)*, pages 111–122, Hague, Netherlands. IOS Press.

[Rodosthenous and Michael, 2019a] Rodosthenous, C. and Michael, L. (2019a). Using Generic Ontologies to Infer the Geographic Focus of Text. In van den Herik, J. and Rocha, A. P., editors, *Agents and Artificial Intelligence*, pages 223–246, Cham. Springer International Publishing.

[Rodosthenous and Michael, 2014] Rodosthenous, C. T. and Michael, L. (2014). Gathering Background Knowledge for Story Understanding through Crowdsourcing. In *Proceedings of the 5th Workshop on Computational Models of Narrative (CMN 2014)*, volume 41, pages 154–163, Quebec, Canada. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[Rodosthenous and Michael, 2018a] Rodosthenous, C. T. and Michael, L. (2018a). A Platform for Commonsense Knowledge Acquisition Using Crowdsourcing. In *Supplementary Proceedings of the enetCollect WG3 & WG5 Meeting 2018*, pages 24–25, Leiden, Netherlands. CEUR.

[Rodosthenous and Michael, 2018b] Rodosthenous, C. T. and Michael, L. (2018b). GeoMantis: Inferring the Geographic Focus of Text using Knowledge Bases. In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*, pages 111–121, Madeira, Portugal. SciTePress.

[Rodosthenous and Michael, 2019b] Rodosthenous, C. T. and Michael, L. (2019b). Web-STAR: A Visual Web-based IDE for a Story Comprehension System. *Theory and Practice of Logic Programming*, 19(2):317–359.

[Singh et al., 2002] Singh, P., Lin, T., Mueller, E., Lim, G., Perkins, T., and Zhu, W. (2002). Open Mind Common Sense: Knowledge acquisition from the general public. *On the*

Move to Meaningful Internet Systems, 2002-DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002, (Davis 1990):1223–1237.

[Speer et al., 2017] Speer, R., Chin, J., and Havasi, C. (2017). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI-17)*, San Francisco, California.

[von Ahn and Dabbish, 2008] von Ahn, L. and Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, 51(8):57.

[Winston, 2012] Winston, P. H. (2012). The Right Way. *Advances in Cognitive Systems*, 1:23–36.

On the Expressive Power of Argumentation Formalisms

Samy Sá

Universidade Federal do Ceará

Abstract

We describe ongoing research on the relative expressive power of argumentation formalisms. Our discussion includes insights and general results on the connections between logic programming, abstract argumentation, assumption-based argumentation and abstract dialectical frameworks. Due to the scope of the paper, instead of formally introducing each system, we offer an abstract set of logical concepts (language, interpretations, models, etc.) pertaining to all argumentation systems. These concepts are used to describe the methodology employed in this type of comparative work and some of the challenges involved. We also discuss open questions and ideas for future research.

1 Introduction

The raise of research on computational argumentation is commonly attributed to [Dung, 1995a]. His work comprises an abstract concept of argument-based reasoning stemming from previous works on abductive logic programming [Eshghi and Kowalski, 1989, Dung, 1995b]. Intuitively, an argument is a sentence in any given logics, therefore sets of arguments are expected to satisfy some criteria of consistency. In the abstract argumentation frameworks of Dung, this criteria is called *admissibility* and sets of arguments satisfying it are called *extensions*. This terminology was employed in several extensions of abstract argumentation frameworks such as assumption-based argumentation [Bondarenko et al., 1997, Dung et al., 2009] and abstract dialectical frameworks [Brewka and Woltran, 2010], but is also found in the literature of abductive reasoning ([Dung, 1995b], for instance). Just alike, virtually the same concepts are found with different terminology in the literature of logic programming semantics [Lloyd, 1987, Przymusinski, 1990] and other non-monotonic reasoning formalisms. The similarities amongst those systems are largely recognized in the literature: each time a new system is proposed, it shows that some of the older systems and their semantics are particular cases of the new. **But just how different these approaches actually are?** That is the main question leading our research.

Our research aims to further the understanding of semantic and syntactic relations between argumentation formalisms. By doing so, we help creating a road map for the adaptation of available tools (such as interpreters, methods, proven results, etc.) between them. For instance, these connections are often verified in a single direction to argue the new systems do at least as much as their predecessors. But more often than not, by the time a new system is proposed, the older systems have several tools readily available. Could they perhaps be adapted? Despite this potential, it is rarely verified whether a new proposal is subsumed by the previous. On that matter, we seek back-and-forth translations and correspondence results (as we have done, for instance, in [Caminada et al., 2015b, Caminada et al., 2015a, Sá and Alcântara, 2019, Alcântara et al., 2019]) to try and answer whether two systems are equivalent or one subsumes the other. If one of the two directions of comparison is already in the literature, we offer results in the complementary direction. Finding more about these relations is of utmost importance

to avoid redundant efforts in the development of the different argumentation systems and their tools.

In this brief paper, we discuss the most common methodology employed to compare the expressive power of two logics and (broadly) some results already obtained. Here, we opt for a rather abstract presentation of argumentation formalisms using common logics terminology. This is meant to avoid the more specific definitions from each system discussed, since there is a large variety of them. Instead, we speak of languages, sentences, theories, and their evaluation as models. Albeit a different nomenclature than usually adopted in argumentation literature, the concepts we will discuss permeate all argumentation formalisms. For this reason, our presentation suffices to explain the relevant methods and results in enough detail and can be considered a contribution of this paper. We also discuss the limitations of currently known translations between theories in the different formalisms and some lines of future research.

2 Method

The general scope of works comparing the expressive power of different logics revolves around translations between theories in those logics and their respective models. To detail our method, we will require some basic logic terminology that permeates all argumentation formalisms. These concepts will later be used to explain what is sought in translations between the theories in different systems.

2.1 Language and Models

All argumentation formalisms stemming from Dung's work and several non monotonic logics preceding them are based on 3-valued interpretations over some particular set of sentences. Here, we will refer to this set of sentences simply as the *language* of the system (denoted $\mathcal{L}$) and consider the *truth-values* in $\mathcal{V} = \{\mathbf{t}, \mathbf{f}, \mathbf{u}\}$, whose elements stand respectively for true, false and undecided. Then, an *interpretation* of $\mathcal{L}$ can be understood simply as a complete function $I : \mathcal{L} \to \mathcal{V}$ that evaluates elements of $\mathcal{L}$ as true, false or undecided. Each argumentation system has some basic criteria for the consistency of interpretations, the most common being Dung's admissibility criteria [Dung, 1995a]. Following the standard nomenclature of logics, here the interpretations of interest for each system will be simply called *models*.

A *theory* is then a series of syntactic specifications about how sentences interfere in the evaluation of one another. As an example, a theory $\mathcal{T}$ may say that the sentence $A \in \mathcal{L}$ can only be true if the sentence $B \in \mathcal{L}$ is not. Commonly, a theory will have many models (because this is only a basic consistency check), but some of those models will have nicer properties than others. Several interesting patterns concerning those nicer properties can be observed in all argumentation systems we studied. The most important pattern is likely the existence of a reference semantics we will here call the *complete* semantics. Based on the complete semantics, a wide range of other semantics can be defined simply by maximizing or minimizing the sets of sentences according to one of the truth values. Therefore, we can generally speak of maximally true complete models, minimally true complete models, and so on (see [Caminada et al., 2015b], for instance), in each argumentation system. Those semantics may receive different names in different nonmonotonic reasoning formalisms, so we will refrain from using their more specific names and simply speak of min/max $\mathbf{t}/\mathbf{f}/\mathbf{u}$ (complete) semantics in each case.

2.2 Translations and Equivalence

Given two argumentation systems S_1, S_2, to find which one is the most expressive, one should investigate how the theories from one system would be modeled in the other. Let $\texttt{Theories}(S_1), \texttt{Theories}(S_2)$ be respectively the sets of all theories that can be proposed in S_1, S_2, a translation from S_1 to S_2 is a total function $\texttt{S}_1\texttt{toS}_2 : \texttt{Theories}(S_1) \to \texttt{Theories}(S_2)$. To ensure an adequate translation, several steps of planning take place. First of all, for each semantics σ of S_1, a corresponding semantics $corr(\sigma)$ of S_2 must be identified. A necessary condition at this point is that if the semantics τ of S_1 can be obtained from σ by a particular operation $f(\sigma) = \tau$, then it must be the case that $f(corr(\sigma)) = corr(\tau)$. There-

fore, one should seek a translation where given $\mathcal{T} \in \texttt{Theories}(S_1)$, there is a $corr(\sigma)$-model of $\texttt{S}_1\texttt{toS}_2(\mathcal{T})$ for each model of $\mathcal{T}$ in σ. This means that a suitable translation rewrites $\mathcal{T}$, originally written in the syntax of S_1, into the syntax of S_2, but in a particular way that the original semantics of $\mathcal{T}$ is preserved. If this much is achieve by $\texttt{S}_1\texttt{toS}_2$, we will have that S_2 *subsumes* S_1 on σ. If the same can be proved for all semantics of S_1, we will (respectively) have that S_2 is *at least as expressive* as S_1. Further, if $\texttt{S}_1\texttt{toS}_2$ is also a bijective function preserving the semantics σ for all S_1-theories, we will have that S_1, S_2 are *equivalent* under σ. The two systems are logically equivalent if a bijective translation is available that preserves all semantics from both systems in one another.

3 Discussion

The methodology here described has been successfully employed in several works to contrast the semantics of abstract argumentation and logic programming [Wu et al., 2009, Caminada et al., 2015b], abstract argumentation and assumption-based argumentation [Caminada et al., 2015a], logic programming and assumption-based argumentation [Caminada and Schulz, 2017], different types of semantics for assumption-based argumentation [Schulz and Toni, 2017, Sá and Alcântara, 2019], different perspectives in the semantics of abstract dialectical frameworks [Alcântara and Sá, 2018], and the semantics of logic programming to abstract dialectical frameworks [Alcântara et al., 2019], amongst other works. As part of this PhD, we made significant contributions in [Caminada et al., 2015b, Caminada et al., 2015a, Alcântara and Sá, 2018, Sá and Alcântara, 2019, Alcântara et al., 2019].

The relative expressive power of argumentation systems is commonly investigated based on their respective complete semantics first. This choice is motivated by some desirable properties of the complete semantics, but also because it generalizes several other semantics in most systems. Commonly, these works found back-and-forth correspondences in all particular cases of the complete semantics, except for the minimally undecided complete semantics.

For instance, when comparing abstract argumentation and logic programming [Wu et al., 2009, Caminada et al., 2015b] the translation from an abstract argumentation framework $\mathcal{F}$ into a logic program $\texttt{AAtLP}(\mathcal{F})$ is very straightforward: there will be an atom in the language of $\texttt{AAtLP}(\mathcal{F})$ for each argument in $\mathcal{F}$ and the rules in the program describe the attack relation from $\mathcal{F}$. This translation ensures there is precisely one sentence in the language of the goal theory (a logic program) for each sentence in the input theory (an argumentation framework). This property makes it drastically simple to prove that logic programming subsumes abstract argumentation in the complete and in all other semantics derived from it. Going the other way around is a bit harder: for each possible derivation (a proof) that is possible in a logic program $\mathcal{P}$, an argument is instantiated in $\texttt{LPtAA}(\mathcal{P})$, then an attack relation is established based on which derivations are supported by each model of the program. In this other translation, there may be multiple arguments in $\texttt{LPtAA}(\mathcal{P})$ for each atom in $\mathcal{P}$, so a bijection between the two languages cannot be ensured. This makes it harder to prove the correspondence in each semantics and also causes an exception involving the semantics that minimize undecided sentences in each system. That is, the complete semantics that minimizes undecided arguments in $\texttt{LPtAA}(\mathcal{P})$ fails to capture the corresponding semantics from logic programming for $\mathcal{P}$. These results propose that logic programming is more expressive than abstract argumentation.

Similar results suggest that assumption-based argumentation is more expressive than abstract argumentation [Caminada et al., 2015a] and logic programming [Bondarenko et al., 1997, Caminada and Schulz, 2017], and that logic programming is more expressive than abstract dialectical frameworks [Alcântara et al., 2019], amongst others. In each case, there is an exception concerning the respective semantics that minimize undecided sentences in the corresponding complete semantics in one of the two directions. Those results are still tied to the translations employed, so there might be alternatives ensuring a correspondence between those systems in all semantics. Fortunately (or not), the accumulated evidence suggests that it is

not the case. Instead, there would be some slight differences that can only be detected by semantics minimizing undecided sentences in each system. This is very curious, as it suggests that we deal with slightly different flavors of undecidability in those systems. In the bigger picture, the results ensure those systems are equivalent in all other cases of the complete semantics, including the more mainstream semantics: stable and grounded (or well-founded). That is, logic programming, abstract argumentation, assumption-based argumentation and abstract dialectical framework may entirely replace one another in several situations and the tools available for each system are compatible with all other systems.

The final steps planned for this thesis regard new perspective in the comparison of logic programming and assumption based argumentation using the semantics proposed in [Sá and Alcântara, 2019]. We will also add results on new translations between abstract argumentation and logic programming designed to close the existing gap concerning their respective semantics that minimize undecidedness.

4 Conclusion

In this paper, we discussed about the relative expressive power of argumentation systems and other nonmonotonic reasoning formalisms. We introduced the important concepts for this line of research using a general terminology that is standard in logics (truth-values, language, interpretations, models, theories) and the methodology commonly employed in these works. This methodology was employed in several works to find numerous correspondences between semantics of argumentation systems. Unfortunately, discussing all related results and works is beyond the scope and length of this paper, so we restricted our attention to some of the more recent works, including our own. As we gather the literature on the subject, we find several argumentation systems are nearly equivalent, with only a marginal exception regarding the minimization of undecided sentences. The evidence supports that these systems can replace one another (concerning expressive power and semantics) and that all tools available to each one may be easily ported to the others.

Several ramifications of our work allow future lines of research (beyond the scope of this PhD), the most trivial being the investigation of other argumentation systems in connection to those we studied. In particular, establishing an equivalence under the complete semantics of another system to either one we mentioned suffices to ensure its equivalence to all the others. A less trivial and potentially more interesting issue regards the general concept of model, which is far more relaxed than the complete models and varies widely in those different systems. Each formalism has a different criteria of basic consistency, but despite similarities, they usually do not correspond. We conjecture that these concepts actually correspond for all systems we mentioned and investigating this matter will likely result in strong intuitions about the different types of undecidability that apparent exist within nonmonotonic reasoning. Alike, it is desirable to further understand the differences between those semantics minimizing undecidedness. If one can establish operations, new translations or even specific criteria to close the gaps on the missing semantics between those formalisms, we may be able to tell at once if they are different and, if that is the case, why. Finally, there is a question of computational complexity. If we were to assume all these argumentation systems are semantically equivalent, the differences would be entirely syntactic. Then we should expect that the syntax of some systems favors more efficient computation of models. At the same time that we can focus on translations between systems to compute their semantics, a better approach would be for the community to improve their systems and compete for computational efficiency in the computation of answers (their models) or simply drop those systems that are proven inefficient to focus on the more prominent ones. This matter should be of high value to the argumentation community for the sake of human resources: if we are developing the same tools under only different syntax, many efforts will be redundant and therefore wasteful; but if we join forces in the development of tools for fewer prominent systems, we encourage cooperation and will likely achieve better results in the long run.

References

[Alcântara and Sá, 2018] Alcântara, J. F. L. and Sá, S. (2018). On three-valued acceptance conditions of abstract dialectical frameworks. In Accattoli, B. and Olarte, C., editors, *Proceedings of the 13th Workshop on Logical and Semantic Frameworks with Applications, LSFA 2018, Fortaleza, Brazil, September 26-28, 2018*, volume 344 of *Electronic Notes in Theoretical Computer Science*, pages 3–23. Elsevier.

[Alcântara et al., 2019] Alcântara, J. F. L., Sá, S., and Guadarrama, J. C. A. (2019). On the equivalence between abstract dialectical frameworks and logic programs. *Theory Pract. Log. Program.*, 19(5-6):941–956.

[Bondarenko et al., 1997] Bondarenko, A., Dung, P., Kowalski, R., and Toni, F. (1997). An abstract, argumentation-theoretic approach to default reasoning. *AI*, 93:63–101.

[Brewka and Woltran, 2010] Brewka, G. and Woltran, S. (2010). Abstract dialectical frameworks. In *Twelfth International Conference on the Principles of Knowledge Representation and Reasoning*, pages 102–111. AAAI Press.

[Caminada et al., 2015a] Caminada, M., Sá, S., Alcântara, J., and Dvorák, W. (2015a). On the difference between assumption-based argumentation and abstract argumentation. *IfCoLog Journal of Logics and their Applications*, 2:15–34.

[Caminada et al., 2015b] Caminada, M., Sá, S., Alcântara, J., and Dvorák, W. (2015b). On the equivalence between logic programming semantics and argumentation semantics. *Int. J. Approx. Reasoning*, 58:87–111.

[Caminada and Schulz, 2017] Caminada, M. and Schulz, C. (2017). On the equivalence between assumption-based argumentation and logic programming. *Journal of Artificial Intelligence Research*, 60:779–825.

[Dung, 1995a] Dung, P. (1995a). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357.

[Dung et al., 2009] Dung, P., Kowalski, R., and Toni, F. (2009). Assumption-based argumentation. In Simari, G. and Rahwan, I., editors, *Argumentation in Artificial Intelligence*, pages 199–218. Springer US.

[Dung, 1995b] Dung, P. M. (1995b). An argumentation-theoretic foundation for logic programming. *The Journal of logic programming*, 22(2):151–177.

[Eshghi and Kowalski, 1989] Eshghi, K. and Kowalski, R. A. (1989). Abduction compared with negation by failure. In Levi, G. and Martelli, M., editors, *Logic Programming, Proceedings of the Sixth International Conference, Lisbon, Portugal, June 19-23, 1989*, pages 234–254. MIT Press.

[Lloyd, 1987] Lloyd, J. W. (1987). *Foundations of Logic Programming; (2nd extended ed.)*. Springer-Verlag New York, Inc., New York, NY, USA.

[Przymusinski, 1990] Przymusinski, T. (1990). The well-founded semantics coincides with the three-valued stable semantics. *Fundamenta Informaticae*, 13(4):445–463.

[Sá and Alcântara, 2019] Sá, S. and Alcântara, J. a. (2019). Interpretations and models for assumption-based argumentation. In *Proceedings of the 34rd Annual ACM Symposium on Applied Computing*, SAC '19, New York, NY, USA. ACM.

[Schulz and Toni, 2017] Schulz, C. and Toni, F. (2017). Labellings for assumption-based and abstract argumentation. *International Journal of Approximate Reasoning*, 84:110–149.

[Wu et al., 2009] Wu, Y., Caminada, M., and Gabbay, D. (2009). Complete extensions in argumentation coincide with 3-valued stable models in logic programming. *Studia Logica*, 93(1-2):383–403. Special issue: new ideas in argumentation theory.

Argumentation-based Dialogue Games for Modelling Deception

Ștefan Sarkadi

Department of Informatics, King's College London, UK

Abstract

Machines of the future might either be endowed with or might develop mechanisms to argue with other agents. We consider the contexts in which these types of machines also develop reasons to act dishonestly by attempting to deceive their interlocutors. Using the argumentation dialogue games approach, this work aims to explore how deceptive machines might be engineered in order to mitigate or neutralise their malicious behaviour. Argumentation dialogue games can be a powerful approach for the modelling of deception given that it offers an *explainable* way of representing the components necessary for deception such as the knowledge of the agents, their ability to perform actions (to communicate arguments), their ability to reason defeasibly about the world, and most importantly, their ability to reason defeasibly about each others' minds. This paper presents three different hybrid agent-based models derived from argumentation that (i) have been successfully used and that (ii) can be used in future work to model machine deception.

1 Introduction

The ability of machines to deceive autonomously is increasingly drawing the interest of the AI community, as well as the interest of the philosophical and digital humanities communities. This has also been enhanced by the emergence of the *post-truth* technological era driven by the popularisation of the term *fake news* [Lazer et al., 2018].

Currently, most of the approaches in AI merely focus on using machine learning capabilities to either (i) generate fake information [Yao et al., 2017], or (ii) detect fake news from big data [Conroy et al., 2015], or (iii) by enhancing machine learning using techniques such as argument mining to detect deception [Cocarascu and Toni, 2016]. However, as pointed out in [Sarkadi, 2018] and in [Sarkadi, 2020b], these data-oriented approaches fail to account for several critical components of deception, namely the intention of the agents to deceive, their *Theory of Mind* of their targets, and the reasoning behind their deceptive acts. Apart from these explainability issues (see [Miller, 2018]) which machine learning approaches face, there is also the issue of representational and design accuracy. Deception and deception detection require *social intelligence*. That is the ability of agents to reason about other minds in order to influence other agents through communication [Castelfranchi, 1998]. In this context, communication represents social actions which influence belief changes, and by extension behavioural changes, in other social agents. Is the AI used behind fake news and deepfakes truly autonomous? Obviously it is not, since it is unable to act autonomously, as well as unable to reason about what information should be fed to whom in order to deceive. These so called types of "deceptive" AI merely act as tools in the hands of truly socially intelligent agents, namely the humans that have the intent and the communicative capabilities to act dishonestly. How do we then model socially-intelligent autonomous deceptive agents that truly behave dishonestly?

In this paper we present a novel argumentation-based dialogue game method to model and study autonomous deceptive agents. This method is extensively described in [Sarkadi, 2020a] and aims to address the problem of modelling deceptive machines from the socially-intelligent agent-based perspective.

2 Method

To model deception, we mainly focus on interactions between two agents, the *Deceiver*, and its target, which we have called the *Interrogator*. The aim of the Deceiver is, obviously, to deceive the Interrogator, whereas the aim of the Interrogator is to find out the desired truth. We have adopted the following definition of deception:

Deception *The intention of a deceptive agent, to make or cause another target agent to believe something is true that the deceiver believes is false, with the aim of achieving an ulterior goal or desire.*

The method we have used to address deception as defined above is the application of opponent modelling to dialogue argumentation games. We have used *belief-desire-intention* BDI-like architectures to model the cognitive properties of the agents that play these dialogue games [McBurney and Parsons, 2009]. Giving BDI agents a communication protocol along with a reasoning mechanism enables them to think pragmatically about their beliefs, their desires, and their intentions in order to perform speech acts [Rao et al., 1995]. In argumentation, these speech acts can represent arguments, as well as argument chains and argument systems. Apart form performing speech acts, our agents are able to reason about their opponent's mind. In other words, they have a Theory of Mind (ToM) that enables mind-reading. ToM is the ability of an agent to reason about the mental attitudes (beliefs, desires, and intentions) of another agent [Goldman, 2012]. According to [Isaac and Bridewell, 2017], mind-reading is a crucial ability for a machine to have in order to be able to deceive or detect deception.

3 Discussion

In this section we present three models that we have built using the method presented in the previous section. All of these models represent deception according to our adopted definition where the Deceiver aims to achieve an ulterior goal. Once this ulterior goal is achieved, deception becomes successful.

The ulterior goal of the Deceiver can even be as simple as desiring that the target believes something is true, when the Deceiver believes it is false. If there is another ulterior goal, that in order to be met requires deception, then the causation of a false belief becomes the subgoal. This is the case in [Panisson et al., 2018], where we have implemented in an agent-oriented programming language a car-dealing agent that deceives in order to cause its target to buy a car the dealer agent desires to sell. The implemented agent can also decide to lie or bullshit. However, we mention the fact that lying or bullshitting is different from deception as they do not require a ToM to cause a false belief.

Another form of ulterior goal is in the case of interrogation games, where the Deceiver needs to cause the Interrogator into accepting the story that emerges from their dialogue. In [Sarkadi et al., 2019a], we have presented an argumentation dialogue game model for generating credible stories. In this model, both Deceiver and Interrogator use the same Toulmin-like reasoning technique to generate simple or complex arguments (that represent stories or narratives) using the ToM of their opponent. The ToM of the opponent contains simple arguments, as well as argument attacks and argument backings, that the opponent knows.

Estimating the success of deception given the communication of an argument can be problematic. The dynamics of deceptive attempts can be influenced by the uncertainty of certain social factors such as: the trust of the interlocutor, the degree of confidence in one's own ToM of the interlocutor, and one's degree of communicative skill. In [Sarkadi et al., 2019b] we have continued our work from [Panisson et al., 2018]. We have

used BDI architectures to model, evaluate and implement in an Agent-oriented Programming Language deceptive interactions under factors of social uncertainty. This model aims to integrate components of two major theories of deception, namely *Interpersonal Deception Theory* [Buller and Burgoon, 1996] and *Information Manipulation Theory 2* [McCornack et al., 2014]. By modelling information manipulation as well as uncertainty of the social factors, we have enabled the Interrogator to consider its own degree of trust in the Deceiver in order to reason about what is being communicated. Given the levels of trust of the Interrogator, we have also enabled the Deceiver to estimate its success at deception by taking into account the trust of its target, the uncertainty of its ToM of the target, as well as its communicative skill. A critical result from the evaluation of the model shows that agents with **strong skeptical attitudes** are prone to **unintended deception**.

Properties

The three models that we have developed in [Panisson et al., 2018], [Sarkadi et al., 2019a], and [Sarkadi et al., 2019b] present different desirable properties that are useful for the study of machine deception (See Table 1). Below we describe each of these properties:

1. **Explainability** should be a crucial property of argumentation-based models of deception. We should be able to evaluate deceptive mind games and say whether deception takes places and if it does we need to explain why and under which conditions it does. An explainable model should be able to inform us if deception can be prevented or mitigated in different contexts.

2. **Unintended Deception** happens when the Deceiver does not attempt deception, but the consequences of its communicative acts result in its potential target to be deceived. It is important for models of deception to be able to represent such unintended consequences as they are critical for accountability. We need to be able to tell if an agent that has the ability to deceive should be held responsible for its actions or not.

3. **Uncertainty** in communication should be considered when modelling deception. This is especially important for modelling an agent that estimates its likelihood of success, as well as modelling agents with different degrees of trust in each other. While most of the times trust should be a default attitude towards others [Levine, 2014], in cases of potential deception this is not the case.

4. **Storytelling** is the ability of an agent to communicate arguments in such a way as to describe to another agent a meaningful chain of events. The ability to build narratives is an emerging topic in AI. Deceptive agents can use this ability to their own benefit, e.g. deliver a fictitious story that compels a jury into absolving them of a crime. Therefore, it should be desirable for models of deception to consider or represent such mechanisms.

5. **Deception Detection** is desirable to be represented in a model. While some models represent and explain why deception is successful, they do not represent how and why deception might be detected. It is also important to distinguish between an agent that has the ability to detect deception and one that has the tendency to believe or not what a Deceiver is communicating, which is the case in some models. Representing deception detection could be also useful in showing how a Deceiver might act knowing that its target is able to detect its deceptive intents, as well as how its target might detect them.

6. **Implementation** is to be desired, but not necessary for modelling deception, or any other social phenomenon. However, demonstrating an implementation of the model helps others to use it for studying different multi-agent system setups and scenarios of social interactions. Implementation also improves the **transparency** of a model,

increasing the model's accessibility through its code.

Model Properties	1	2	3	4	5	6
[Panisson et al., 2018]	✓	-	-	-	-	✓
[Sarkadi et al., 2019a]	✓	-	-	✓	✓	-
[Sarkadi et al., 2019b]	✓	✓	✓	-	-	✓

Table 1: Comparison of our models in terms of their respective desirable properties for the study of machine deception.

Future Work

Modelling deception using dialogue games for argumentation offers explanatory and representational power, especially if we want to show properties such as the ones expressed by the models we introduced and described here. However, none of the models presented expresses the full spectrum of these properties. By no means should this discourage the continuation of our method for the study of deception. This paper has classified *a posteriori* the models according to the properties they have managed to express. The models have not been designed and built starting from an *a priori* knowledge of this classification. Having defined this classification should help us continue using our method towards building more expressive models of machine deception.

A problem that future work should aim to overcome is the introduction of an environment in the socio-dynamical representations of deception using dialogue games. Our models have mainly focused on the direct interaction between the Deceiver and its target, but unfortunately they have failed to take into account how a Deceiver might use the environment for manipulating the beliefs of its target. We believe this could be a viable research path worth pursuing in the modelling of machine deception.

4 Conclusion

In this paper we have presented an argumentation-based dialogue game method for modelling deception in AI that is extensively described in [Sarkadi, 2020a]. We have introduced two critical components for representing deception, namely BDI agent architectures and Theory of Mind. Our method relies on these components for representing social interactions between two agents, the Deceiver and the Interrogator. We have described and compared three models that we have built using the presented method. We have also compared the models according to their results and to several desirable properties introduced in the paper, namely *explainability, unintended deception, uncertainty, storytelling, deception detection, and implementation.*

Acknowledgements

The research described by this short paper could not have been possible without their influence and/or contribution: R.H. Bordini, M. Chapman, P. McBurney, F. Mosca, A.R. Panisson and S. Parsons.

References

[Buller and Burgoon, 1996] Buller, D. B. and Burgoon, J. K. (1996). Interpersonal Deception Theory. *Communication Theory*, 6(3):203–242.

[Castelfranchi, 1998] Castelfranchi, C. (1998). Modelling social action for ai agents. *Artificial intelligence*, 103(1-2):157–182.

[Cocarascu and Toni, 2016] Cocarascu, O. and Toni, F. (2016). Detecting deceptive reviews using argumentation. In *Proceedings of the 1st International Workshop on AI for Privacy and Security*, page 9. ACM.

[Conroy et al., 2015] Conroy, N. J., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

[Goldman, 2012] Goldman, A. I. (2012). Theory of mind. In *The Oxford Handbook of Philosophy of*

Cognitive Science, volume 1. Oxford Handbooks Online, 2012 edition.

[Isaac and Bridewell, 2017] Isaac, A. and Bridewell, W. (2017). *White lies on silver tongues: Why robots need to deceive (and how)*. Oxford University Press.

[Lazer et al., 2018] Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

[Levine, 2014] Levine, T. R. (2014). Truth-Default Theory (TDT). *Journal of Language and Social Psychology*, 33(4):378–392.

[McBurney and Parsons, 2009] McBurney, P. and Parsons, S. (2009). Dialogue games for agent argumentation. In Simari, G. and Rahwan, I., editors, *Argumentation in Artificial Intelligence*, pages 261–280. Springer US.

[McCornack et al., 2014] McCornack, S. A., Morrison, K., Paik, J. E., Wisner, A. M., and Zhu, X. (2014). Information manipulation theory 2: a propositional theory of deceptive discourse production. *Journal of Language and Social Psychology*, 33(4):348–377.

[Miller, 2018] Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.

[Panisson et al., 2018] Panisson, A. R., Sarkadi, S., McBurney, P., Parsons, S., and Bordini, R. H. (2018). Lies, bullshit, and deception in agent-oriented programming languages. In *20th International Trust Workshop (co-located with AAMAS/IJCAI/ECAI/ICML 2018), Stockholm, Sweden, 14 July, 2018*, pages 50–61. CEUR-WS.

[Rao et al., 1995] Rao, A. S., Georgeff, M. P., et al. (1995). BDI agents: from theory to practice. In *ICMAS*, volume 95, pages 312–319.

[Sarkadi, 2018] Sarkadi, S. (2018). Deception. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5781–5782. AAAI Press.

[Sarkadi, 2020a] Sarkadi, S. (2020a). *Deception*. PhD thesis, King's College London.

[Sarkadi, 2020b] Sarkadi, S. (2020b). Deceptive autonomous agents. Cranfield Online Research Data (CORD).

[Sarkadi et al., 2019a] Sarkadi, S., McBurney, P., and Parsons, S. (2019a). Deceptive storytelling in artificial dialogue games. In *Proceedings of the 2019 AAAI Spring Symposium Series on Story-Enabled Intelligence*.

[Sarkadi et al., 2019b] Sarkadi, S., Panisson, A., Bordini, R., McBurney, P., Parsons, S., and Chapman, M. (2019b). Modelling deception using theory of mind in multi-agent systems. *AI COMMUNICATIONS*.

[Yao et al., 2017] Yao, Y., Viswanath, B., Cryan, J., Zheng, H., and Zhao, B. Y. (2017). Automated crowdturfing attacks and defenses in online review systems. *arXiv preprint arXiv:1708.08151*.

On the Link Between Truth Discovery and Bipolar Abstract Argumentation

Joseph Singleton

School of Computer Science and Informatics, Cardiff University

Abstract

With the abundance of data available in today's world, e.g. from the web, social media platforms and crowdsourcing systems, it is common to find conflicting information from different data sources. *Truth discovery* algorithms aim to find the true 'facts' amongst such conflicts by estimating the trustworthiness of data sources, so that the claims made by trustworthy sources can be given priority. Since source trustworthiness is unknown *a priori*, such algorithms jointly estimate trust in sources and belief in facts, assigning higher trust scores to sources who claim believable facts and higher belief scores to facts claimed by trusted sources. Truth discovery has received increasing attention in the data mining and crowdsourcing literature, but other perspectives may offer additional insight into the problem and potential solutions. In this paper we discuss the link between truth discovery and argumentation, with a particular focus on *bipolar abstract argumentation*.

1 Introduction

Information is available in ever-increasing quantities in today's world. The web, social media platforms and crowdsourcing systems host vast amounts of data from a diverse range of sources, including individual users, websites and smartphone sensors. Sources naturally vary in their reliability and the quality of data they produce, which can cause conflicts when multiple sources comment on the same issue. This poses a problem for automatically extracting information on factual matters: which sources should we trust, and which 'facts' should we believe?

Truth discovery algorithms aim to solve this issue by estimating the *trustworthiness* of data sources [Li et al., 2016]. Facts claimed by trustworthy sources are given high weight in producing the aggregated output, whereas those claimed by untrustworthy sources have little impact. Truth discovery has been studied in the data mining and crowdsourcing literature, where unsupervised algorithms jointly estimate the trustworthiness of sources and a measure of belief in the facts being proposed. However, other perspectives may offer additional insight into the problem and its potential solutions. In this paper we take preliminary steps to explore *argumentation*-based approaches to truth discovery.

Argumentation deals with finding sets of 'acceptable' arguments given the attacks (i.e. conflicts) between them. This shares a high-level similarity with truth discovery, where we aim to produce coherent outputs by resolving conflicts between claims from multiple sources. One may therefore wonder whether truth discovery can be rephrased as an argumentation problem. If so, argumentation semantics would translate to new algorithms for truth discovery. Such algorithms would make explicit the indirect conflicts and attacks present in truth discovery, and potentially lead to explainable algorithms (e.g. via discussion games).

Many systems of argumentation have been studied in the literature and could be applied to truth discovery. In the present paper we narrow our focus to *bipolar abstract argumentation* [Cayrol and Lagasquie-Schiex, 2005] – where there

are both attack and support relations between arguments – and give a simple method for converting a truth discovery problem into one of bipolar argumentation.

The paper is structured as follows. In section 2 we formally define the truth discovery problem. Section 3 briefly reviews different approaches to abstract argumentation from the literature. Section 4 shows how truth discovery may be formulated in terms of bipolar argumentation, and we conclude in section 5.

2 Truth Discovery Problem Formulation

To provide context for the rest of the discussion, we formally state the truth discovery problem in this section. This formulation is a minor variation of our previous work [Singleton and Booth, 2020].

We consider fixed disjoint sets $\mathcal{S}, \mathcal{O}$ and $\mathcal{F}$ which represent *sources*, *objects* and *facts* respectively. A source represents an entity which provides data, e.g. a website, an individual or piece of sensor equipment. An object represents a real-world entity or question, e.g. 'What is the height of Mount Everest?'. A fact represents a piece of information, relating to an object, which a source may claim is true, e.g. 'Mount Everest is 8,850m tall'. An object may have several conflicting facts relating to it; one of the aims of truth discovery is to determine which fact should be believed.

The input to truth discovery, which we term a *truth discovery network*, consists of a set of claims between sources and facts, and the links between facts and objects. We add the constraints that each fact is related to a single object, and sources cannot claim conflicting facts.

Definition 1. *A* truth discovery network *(hereafter a* TD network*) is a pair $N = \langle \mathcal{C}, \mathcal{L} \rangle$ where*

1. *$\mathcal{C} \subseteq \mathcal{S} \times \mathcal{F}$ and $\mathcal{C} \neq \emptyset$*
2. *$\mathcal{L} \subseteq \mathcal{F} \times \mathcal{O}$ and $\mathcal{L} \neq \emptyset$*
3. *For each $f \in \mathcal{F}$ there is a unique $o \in \mathcal{O}$ such that $(f, o) \in \mathcal{L}$. We denote such o by $\mathsf{obj}_N(f)$*
4. *For each $s \in \mathcal{S}$ and $f, g \in \mathcal{F}$, if $(s, f), (s, g) \in \mathcal{C}$ then $\mathsf{obj}_N(f) \neq \mathsf{obj}_N(g)$*

Let $\mathcal{N}$ denote the set of all TD networks.

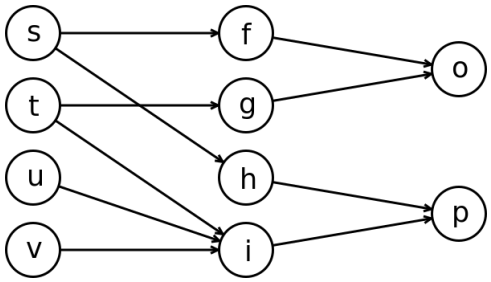


Figure 1: Example graph representation of a truth discovery network, with sources s, t, u, v, objects o, p and facts f, g, h, i. Here s and t disagree on the true fact for objects o and p. Sources u and v do not comment on object o, but agree with t on object p.

Example 1. Consider the TD network $N = \langle \mathcal{C}, \mathcal{L} \rangle$ where $\mathcal{C} = \{(s, f), (t, g), (s, h), (t, i), (u, i), (v, i)\}$ and $\mathcal{L} = \{(f, o), (g, o), (h, p), (i, p)\}$. This network can be visualised as a directed graph whose nodes are $\mathcal{S} \cup \mathcal{F} \cup \mathcal{O}$, and whose edges are $\mathcal{C} \cup \mathcal{L}$, as shown in fig. 1.

The output of truth discovery is an assessment of the trustworthiness of sources and the believability of facts. We also allow *multiple* outputs, in analogy with how argumentation semantics may yield multiple extensions.

Definition 2. *A* truth discovery operator *(hereafter a* TD operator*) is a mapping $T : \mathcal{N} \to \mathbb{R}^{\mathcal{S} \cup \mathcal{F}}$. We write T_N for $T(N)$ so that $T_N : \mathcal{S} \cup \mathcal{F} \to \mathbb{R}$ is a mapping which assigns each source a* trust score *and each fact a* belief score*. A* multi-valued operator *is a mapping $T : \mathcal{N} \to \mathcal{P}(\mathbb{R}^{\mathcal{S} \cup \mathcal{F}})$, where $\mathcal{P}$ denotes the power set operation. That is, T_N is a set of mappings $\mathcal{S} \cup \mathcal{F} \to \mathbb{R}$.*

Note that a (single valued) operator induces total preorders on $\mathcal{S}$ and $\mathcal{F}$ for each network N – these represent the *trust* and *belief* rankings in sources and facts respectively. That is, $T_N(s_1) \leq T_N(s_2)$ means source s_2 is considered at least as trustworthy as s_1, and $T_N(f_1) \leq T_N(f_2)$ means that fact f_2 is at least as believable as f_1.

3 Argumentation Frameworks

Many kinds of abstract argumentation frameworks (AFs) have been studied in the literature. At a minimum, an AF consists of a set of arguments and some interactions between them. Arguably the most well-known is Dung's *abstract argumentation framework*, which is a pair $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ with $\mathcal{A}$ being a set of arguments and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ the *attack relation* [Dung, 1995].

Here arguments interact only through attacks. Whilst a notion of *defense* exists by combining two chained attacks, it has been argued that this is not sufficient for all examples, and an independent notion of *support* is required [Cayrol and Lagasquie-Schiex, 2005]. This leads to the definition of a *bipolar argumentation framework*.

Definition 3 ([Cayrol and Lagasquie-Schiex, 2005])**.** *An* abstract bipolar argumentation framework *(BAF) is a tuple* $\langle \mathcal{A}, \mathcal{R}_{att}, \mathcal{R}_{sup} \rangle$ *where* $\mathcal{A}$ *is a set of arguments,* $\mathcal{R}_{att} \subseteq \mathcal{A} \times \mathcal{A}$ *is the* attack relation*, and* $\mathcal{R}_{sup} \subseteq \mathcal{A} \times \mathcal{A}$ *is the* support relation*.*

The addition of a support relation is appealing from the truth discovery point of view, since an important component of a TD network is the *support* between sources and the facts they claim are true.

Other work in the literature extends Dung's model in other ways to encode additional information. Many such extensions are appropriate for handling multi-agent argumentation and reasoning about trust and reliability in arguments. For example, social abstract argumentation [Leite and Martins, 2011] models users who vote for/against arguments before semantics are applied. Weighted argumentation [Baroni et al., 2019] equips arguments with an initial weight, which can be used to represent the reliability of the source putting the argument forward. Similarly, weights may be applied to the attack or support interactions [Janssen et al., 2008]. Trust between multiple agents has also been studied, where each agent constructs an AF based on the beliefs of others and the trust placed in them [Parsons et al., 2011].

We leave the study of these areas in relation to truth discovery for future work, and focus solely on bipolar argumentation in the present paper.

4 Argumentation-based Truth Discovery

In this section we describe a method for translating a TD network into a BAF, and illustrate how bipolar semantics may be applied to truth discovery with an example. It must be stressed that this is not the only possible method for reframing truth discovery in terms of argumentation; nor is it the only method even for bipolar argumentation. Nonetheless, we consider this to be a necessary first step towards exploration of argumentation-based methods.

4.1 Constructing a Bipolar Framework

Given a TD network as per definition 1, we must construct a BAF. This involves defining the arguments and the attack and support relations.

Defining the Arguments. Recall that the outputs of (abstract) argumentation tell us which sets of arguments are acceptable given the interactions between arguments. Now, in truth discovery we are interested in not just which facts to believe, but also which sources to trust. Moreover these two halves of the problem must cohere with one another: a fact claimed by trusted sources should be believed, and a source claiming believable facts should be trusted.

To maintain this symmetry, we propose to encode both source trustworthiness and fact believability through the arguments. That is, for each source s we introduce an argument "s is a trustworthy source", and for each fact f we introduce an argument "f is a believable fact". Identifying such arguments with the sources and facts themselves, we take the set of arguments to be $\mathcal{A} = \mathcal{S} \cup \mathcal{F}$. Note that this is contrary to other approaches to multi-source argumentation, which consider the sources to exist at a level above the arguments (e.g. [Parsons et al., 2011]). Note also that the objects $\mathcal{O}$ are not explicitly represented as arguments, and instead play a role in the construction of attacks below.

Defining the Attack and Support Relations. The inherent conflicts in a TD network lie between mutually exclusive facts that relate to the same object. We therefore introduce an attack between (the

arguments corresponding to) facts f and g whenever $\mathsf{obj}_N(f) = \mathsf{obj}_N(g)$ and $f \neq g$. Note that this may be expressed by a 'mutual exclusion' relation $\mathcal{M} = (\mathcal{L} \circ \mathcal{L}^{-1}) \setminus i_{\mathcal{F}}$, which is symmetric.[1]

When it comes to the support relation, an intuitive definition is that sources support the believability of the facts they claim, and facts support the trustworthiness of the sources claiming them.[2] That is, the support relation is $\mathcal{R}_{\text{sup}} = \mathcal{C} \cup \mathcal{C}^{-1}$. In full, our translation from TD networks to BAFs is as follows.

Definition 4. *The BAF associated with a network* $N = \langle \mathcal{C}, \mathcal{L} \rangle$ *is* $B(N) = \langle \mathcal{A}, \mathcal{R}_{att}, \mathcal{R}_{sup} \rangle$ *where* $\mathcal{A} = \mathcal{S} \cup \mathcal{F}$, $\mathcal{R}_{att} = \mathcal{M} = (\mathcal{L} \circ \mathcal{L}^{-1}) \setminus i_{\mathcal{F}}$ *and* $\mathcal{R}_{sup} = \mathcal{C} \cup \mathcal{C}^{-1}$.

Note that while attacks only exist between facts, indirect conflicts in the TD network can be expressed through *complex attacks* [Cayrol and Lagasquie-Schiex, 2013]; e.g. sources have a *supported attack* on facts conflicting with their beliefs and a *super-mediated attack* on sources with whom they disagree.

4.2 Argumentation Semantics

Equipped with a mapping B from TD networks to BAFs, we can form a TD operator by applying semantics to $B(N)$. Identifying which semantics yield intuitive results for truth discovery is an interesting task for future research. Here we aim only to illustrate the idea with a simple example; for this we take the meta-argumentation approach of [Cayrol and Lagasquie-Schiex, 2013]. This method uses the attack and support relations to form new 'meta-arguments' consisting of *sets* of arguments from the BAF. Applying Dung's complete semantics to the resulting meta AF yields a multi-valued TD operator T^{meta}.[3]

Let N be the TD network from example 1. Deferring the technical details of the meta-argumentation system to the original paper, $B(N)$ yields two meta-arguments X_1, X_2, where $X_1 = \{s, f, h\}$, $X_2 = \{t, u, v, g, i\}$, and each argument attacks the other. The complete extensions are therefore $\{X_1\}$, $\{X_2\}$ and $\emptyset$. Hence $T_N^{\text{meta}} = \{\phi_1, \phi_2, \phi_3\}$ where for $i \in \{1, 2\}$ we have $\phi_i(x) = 1$ if $x \in X_i$ and $\phi_i(x) = 0$ otherwise ($x \in \mathcal{S} \cup \mathcal{F}$), and ϕ_3 is constant 0.

ϕ_1 and ϕ_2 represent two opposite points of view that may be taken in the TD network: one may either trust s and its associated facts, rejecting the conflicting facts and disagreeing sources, or apply the same reasoning with source t. The final possibility ϕ_3, which corresponds to the grounded extension, declines to trust or believe *anything*.

For practical applications of argumentation-based truth discovery, extension-based operators such as T^{meta} may be inappropriate due to their absolutist nature: a source (resp. fact) is either trustworthy (resp. believable) or not, with no middle ground. To take a more fine-grained point of view one may instead apply *gradual* semantics [Baroni et al., 2019], where a numerical *acceptability degree* is assigned to each argument. We leave this to future work.

5 Conclusion

This paper has taken preliminary steps towards truth discovery methods based on bipolar argumentation. However many unanswered questions remain, such as how our TD network to BAF translation compares to other possibilities and which bipolar semantics should be applied. Any new argumentation-based TD operators should also be evaluated – this can be done experimentally or by consideration of which desirable theoretical properties operators have, i.e. *axioms* for truth discovery [Singleton and Booth, 2020]. Such axioms could be compared against similar axioms in gradual argumentation [Amgoud and Ben-Naim, 2016] to see whether properties of semantics translate into properties for truth discovery. More broadly, other argumentation systems besides bipolar should also be investigated to establish their applicability to truth discovery and justify our approach.

Acknowledgements

I thank Richard Booth for our fruitful discussions which formed the basis of this work, and for his in-

[1] Here $\circ$ denotes composition of relations, $\mathcal{L}^{-1}$ denotes the symmetric inverse of $\mathcal{L}$, and $i_{\mathcal{F}}$ denotes the identity relation on $\mathcal{F}$.

[2] This is the intuition underlying many existing TD operators in the literature.

[3] Specifically, we define $T_N^{\text{meta}}(x) = 1$ if the meta-argument corresponding to x is accepted under complete semantics in the meta AF, and $T_N^{\text{meta}}(x) = 0$ otherwise ($x \in \mathcal{S} \cup \mathcal{F}$).

sightful comments of draft versions of the document.

References

[Amgoud and Ben-Naim, 2016] Amgoud, L. and Ben-Naim, J. (2016). Axiomatic foundations of acceptability semantics. In *Proc. KRR*.

[Baroni et al., 2019] Baroni, P., Rago, A., and Toni, F. (2019). From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *International Journal of Approximate Reasoning*, pages 252 – 286.

[Cayrol and Lagasquie-Schiex, 2005] Cayrol, C. and Lagasquie-Schiex, M. C. (2005). On the Acceptability of Arguments in Bipolar Argumentation Frameworks. In *Proc. ECSQARU*, pages 378–389.

[Cayrol and Lagasquie-Schiex, 2013] Cayrol, C. and Lagasquie-Schiex, M. C. (2013). Bipolarity in argumentation graphs: Towards a better understanding. *Int. J. Approx. Reasoning*, pages 876–899.

[Dung, 1995] Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, pages 321 – 357.

[Janssen et al., 2008] Janssen, J., De Cock, M., and Vermeir, D. (2008). Fuzzy argumentation frameworks. In *Proc. IPMU*, pages 513–520.

[Leite and Martins, 2011] Leite, J. and Martins, J. (2011). Social abstract argumentation. In *Proc. IJCAI*, pages 2287–2292.

[Li et al., 2016] Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., and Han, J. (2016). A Survey on Truth Discovery. *SIGKDD Explor. Newsl.*, pages 1–16.

[Parsons et al., 2011] Parsons, S., Tang, Y., Sklar, E., McBurney, P., and Cai, K. (2011). Argumentation-based reasoning in agents with varying degrees of trust. In *Proc. AAMAS*, pages 879–886.

[Singleton and Booth, 2020] Singleton, J. and Booth, R. (2020). An axiomatic approach to truth discovery (extended abstract). In *Proc. AAMAS*. Forthcoming.

A First Idea for a Ranking-Based Semantics using System Z

Kenneth Skiba

Institute for Web Science and Technologies, University of Koblenz-Landau, Germany

Abstract

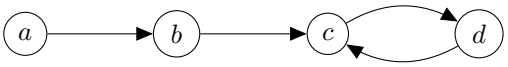
We discuss ranking arguments from an Dung-style argumentation framework with the help of conditional logics. Using an intuitive translation for an argumentation framework to generate conditionals, we can apply nonmonotonic inference systems to generate a ranking on these conditionals. With this ranking we construct a ranking for our arguments.

1 Introduction

Abstract argumentation has become a popular topic in artifical intelligence, mainly in the area of Knowledge Representation and Reasoning. Especially in this area, ranking the solutions does receive increasing attention over the last few years. The usual approach to reason using abstract argumentation frameworks is to differentiate between "accepted" and "rejected" arguments. But using a ranking over the arguments yield to a more fine-grained approach.

There are already ideas for ranking arguments, named ranking-based semantics [Amgoud and Ben-Naim, 2013], like a ranking with respect to a categoriser function [Pu et al., 2014] or based on a two-person zero-sum strategic game [Matt and Toni, 2008] and many more. [Delobelle, 2017] summarizes the state-of-the-art models for ranking arguments. We want to use a novel approach by using conditional logics for our ranking model. Conditional logic is a general non-monotonic representation formalism that focuses on default rule of the form "if A then B" and there exist some interesting relationships between this formalism and that of formal argumentation [Kern-Isberner and Thimm, 2018, Heyninck et al., 2020].

Figure 1: Argumentation framework from Example 1



The rest of this work is organized as follows: In Section 2 all necessary preliminaries will be stated. Then we discus our ranking idea in Section 3 and with Section 4 we conclude this paper.

2 Background

In the following, we want to briefly recall some general preliminaries on conditional logic and argumentation frameworks.

2.1 Abstract Argumentation Frameworks

In this work we use the idea of *argumentation frameworks* first introduced in [Dung, 1995]. An *argumentation framework* AF is a pair $\langle \mathcal{A}, \mathcal{R} \rangle$, where $\mathcal{A}$ is a finite set of arguments and $\mathcal{R}$ is a set of attacks between arguments with $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$. An argument a is said to *attack* b if $(a, b) \in \mathcal{R}$. We call an argument a *acceptable with respect to a set* $S \subseteq \mathcal{A}$ if for each attacker $b \in \mathcal{A}$ of this argument a with $(b, a) \in \mathcal{R}$, there is an argument $c \in S$ which attacks b, i.e., $(c, b) \in \mathcal{R}$; we then say that a is *defended by* c. An argumentation framework $\langle \mathcal{A}, \mathcal{R} \rangle$ can be illustrated by a directed graph with vertex set $\mathcal{A}$ and edge set $\mathcal{R}$.

Example 1. Let $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ with $\mathcal{A} = \{a, b, c, d\}$ and $\mathcal{R} = \{(a, b), (b, c), (c, d), (d, c)\}$ be an argumentation framework. The corresponding graph is shown in Figure 1. Argument b is not acceptable with respect to any set S of arguments, as b is not defended

against a's attack. On the other hand, c is acceptable with respect to $S = \{a, c\}$, as a defends c against b's attack and c defends itself against d's attack.

Up to this point the arguments can only have the two statuses of accepted or not accepted[1], but we want to have a more fine-graded comparison between arguments. For this we use the idea of rankings-based semantics [Amgoud and Ben-Naim, 2013, Delobelle, 2017].

Definition 2 (Ranking-based semantics)**.** A *ranking-based semantics* σ associates to any argumentation framework $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ a ranking $\succeq^{\sigma}_{AF}$ on $\mathcal{A}$ where $\succeq^{\sigma}_{AF}$ is a preorder on $\mathcal{A}$. $a \succeq^{\sigma}_{AF} b$ means that a is at least as acceptable as b. With $a \simeq^{\sigma}_{AF} b$ we describe that a and b are equally acceptable. Finally we say a is strictly more acceptable than b, when $a \succ^{\sigma}_{AF} b$.

We denote by $\sigma(AF)$ the ranking on $\mathcal{A}$ returned by σ.

2.2 Conditional Logics

A *possible world* w for a propositional language is a function, which evaluates an atom a in this language with TRUE or FALSE. We say w satisfies an atom a if $w(a) =$ TRUE, written $w \vdash a$. We denote the set of all interpretations as $\Omega(A)$. We will abbreviate an interpretation w with its *complete conjuction*, i.e., if $a_1, \ldots, a_n \in A$ are the atoms that are assigned TRUE by w and $a_{n+1}, \ldots, a_m \in A$ are the ones assigned with $\bot$, w will be identify $a_1 \ldots a_n \overline{a_{n+1}} \ldots \overline{a_m}$.

As a conditional we interpret a pair $(\varphi|\phi)$ with the information meaning *"if ϕ is* TRUE *then φ is* TRUE *as well"*. For conditional logics we use the approach from [De Finetti, 2017], who considers conditionals as *generalized indicator functions* for possible worlds resp. propositional interpretations w:

$$((\varphi|\phi))(w) = \begin{cases} 1 : w \vdash \phi \wedge \varphi \ (verifies) \\ 0 : w \vdash \phi \wedge \neg\varphi \ (falsifies) \\ u : w \vdash \neg\phi \ (not\ applicable) \end{cases} \quad (1)$$

where u stand for *unknown*. Informal speaking a world w *verifies* a conditional $(\varphi|\phi)$ iff it satisfies both antecedent and conclusion ($(\varphi|\phi)(w) = 1$); it *falsifies* iff it satisfies the antecedence but not the conclusion ($(\varphi|\phi)(w) = 0$); otherwise the conditional is *not applicable* ($(\varphi|\phi)(w) = u$). A conditional $(\varphi|\phi)$ is satisfied by w if it does not falsify it.

Semantics are given to sets of conditionals via ranking functions [Goldszmidt and Pearl, 1996, Spohn, 1988]. With a ranking function, also called *ordinal conditional function (OCF)*, $\kappa : \Omega(A) \to \mathbb{N} \cup \{\infty\}$ we can express the degree of plausibility of possible worlds $\kappa(\phi) := min\{\kappa(w) | w \vdash \phi\}$. With the help of OCFs κ we can express the acceptance of conditionals and nonmonotonic inferences, so $(\varphi|\phi)$ is accepted by κ iff $\kappa(\phi \wedge \varphi) < \kappa(\phi \wedge \varphi)$. With $Bel(\kappa) = \{\phi | \forall w \in \kappa^{-1}(0) : w \vdash \phi\}$ we denote the most plausible worlds.

As there are an infinite number of ranking functions that accept a given set of conditionals, we consider System Z [Goldszmidt and Pearl, 1996] as an inference relation, which yields us a uniquely defined ranking function for reasoning.

Definition 3 (System Z)**.** $(\varphi|\phi)$ is tolerated by a finite set of conditionals Δ if there is a possible world w with $(\phi|\varphi)(w) = 1$ and $(\phi'|\varphi')(w) \neq 0$ for all $(\phi'|\varphi') \in \Delta$. The *Z-partition* $(\Delta_0, \ldots, \Delta_n)$ of Δ is defined as:

- $\Delta_0 = \{\delta \in \Delta | \Delta \ tolerates \ \delta\}$
- $\Delta_1, \ldots, \Delta_n$ is the Z-partition of $\Delta \setminus \Delta_0$

For $\delta \in \Delta$: $Z_\Delta(\delta) = i$ iff $\delta \in \Delta_i$ and $\Delta_1, \ldots, \Delta_n$ is the Z-partitioning of Δ.

We define a ranking function $\kappa^Z_\Delta : \Omega \to \mathbb{N} \cup \{\infty\}$ as $\kappa^Z_\Delta(w) = max\{Z(\delta) | \delta(w) = 0, \delta \in \Delta\} + 1$, with $max\ \emptyset = -1$. Finally $\Delta \mid\sim_Z \phi$ if and only if $\phi \in Bel(\kappa^Z_\Delta)$.

Example 4. Let $\Delta = \{(a|\neg b), (b|\neg a), (c|\neg b \wedge \neg a \wedge \neg d), (d|\top), (c|\neg d)\}$. For this set of conditionals, $\Delta = \Delta_0 \cup \Delta_1$ with $\Delta_0 = \{(a|\neg b), (b|\neg a), (c|\neg b \wedge \neg a \wedge \neg d)\}$ and $\Delta_1 = \{(\neg a \wedge \neg b|\ d)\}$ therefore we have the values from Table 1. So we can derive $(\kappa^Z_{\Delta_0})^{-1}(0) = \{abcd, ab\bar{c}d, a\bar{b}cd, a\bar{b}\bar{c}d, \bar{a}bcd, \bar{a}b\bar{c}d\}$ and $(\kappa^Z_{\Delta_1})^{-1}(0) = \emptyset$.

3 Discussion

In this work we want to extend the work of [Heyninck et al., 2020] and [Kern-Isberner and Thimm, 2018] to not only combine abstract argumentation

[1] Using labeling-based semantics we can generate a three-valued model [Wu et al., 2010].

Table 1: Values for Example 4

ω	$Z((a\|\neg b))$	$Z((b\|\neg a))$	$Z((c\|\neg b \wedge \neg a \wedge \neg d))$	$Z((d\|\top))$	$Z((\neg a \wedge \neg b\|d))$
$abcd$	u	u	u	1	0
$abc\overline{d}$	u	u	u	0	u
$ab\overline{c}d$	u	u	u	1	0
$ab\overline{c}\overline{d}$	u	u	u	0	u
$a\overline{b}cd$	1	u	u	1	0
$a\overline{b}c\overline{d}$	1	u	u	0	u
$a\overline{b}\overline{c}d$	1	u	u	1	0
$a\overline{b}\overline{c}\overline{d}$	1	u	u	0	u
$\overline{a}bcd$	u	1	u	1	0
$\overline{a}bc\overline{d}$	u	1	u	0	u
$\overline{a}b\overline{c}d$	u	1	u	1	0
$\overline{a}b\overline{c}\overline{d}$	u	1	u	0	u
$\overline{a}\overline{b}cd$	0	0	u	1	1
$\overline{a}\overline{b}c\overline{d}$	0	0	1	0	u
$\overline{a}\overline{b}\overline{c}d$	0	0	u	1	1
$\overline{a}\overline{b}\overline{c}\overline{d}$	0	0	0	0	u

and conditional logics, but also present ideas to rank arguments using this combination.

The general idea is to represent an abstract argumentation framework as a set of conditionals, using System Z in order to determine a ranking function that accepts these conditionals, and then extract rankings on arguments from this ranking function. First we need a translation from an argumentation framework to conditional logics. It is clear, that for an argument to be acceptable every attacker has to be not acceptable. With this idea we can construct the conditional logic knowledge base. Let AF be an argumentation framework and $\theta : \mathcal{A} \to \mathcal{C}_{\mathcal{A}}$, where $\mathcal{C}_{\mathcal{A}}$ is the set of conditional knowledge bases over the propositional language generated by $\mathcal{A}$.

$$\theta(AF) = \{(a|B)|(a \in \mathcal{A}), B = \neg b_1 \wedge \neg b_2 \wedge \ldots \wedge \neg b_n, \ where \ (b_i, a) \in \mathcal{R}\} \quad (2)$$

In other words, θ models that an argument is accepted if all its attackers are not accepted.

We can use inference systems like system Z [Goldszmidt and Pearl, 1996] on these conditional knowledge bases to generate a ranking over these worlds. Based on this ranking we want to rank the arguments. A first idea is to count the number of occurrences of a positive literal in the set of worlds $(\kappa_\Delta^Z)^{-1}(0)$ and then rank the corresponding arguments based on this number. So if an argument a has a higher count then an argument b, we say $a \succeq b$. This simple idea yields a clear and uniquely defined ranking, while not needing an complex algorithm to be computed.

Definition 5. Let $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ be an argumentation framework translated with help of $\theta(AF)$ and a inference system to the set of worlds κ_Δ^Z. Define

$$Ccs^{\theta}_{\kappa^Z_\Delta(\omega)}(a) = |\{w \in (\kappa_\Delta^Z)^{-1}(0)|w \vdash a\}| \quad (3)$$

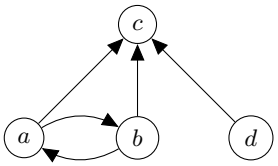
We can then use this counting function for our ranking-based semantics.

Definition 6 (Conditional-counting-based semantics)**.** The *Conditional-counting-based semantics (Ccbs)* associates to any argumentation framework $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ a ranking $\succeq_{AF}^{Ccbs}$ on $\mathcal{A}$ such that $\forall a, b \in \mathcal{A}$ with respect to a transition θ and a ranking function $\kappa_\Delta^Z(\omega)$.

$$a \succeq_{AF}^{Ccbs} b \text{ if and only if } Ccs^{\theta}_{\kappa^Z_\Delta(\omega)}(a) \geq Ccs^{\theta}_{\kappa^Z_\Delta(\omega)}(b)$$

Example 7. Let $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ with $\mathcal{A} = \{a, b, c, d\}$ and $\mathcal{R} = \{(a, b), (b, a), (a, c), (b, c), (d, c)\}$

Figure 2: Argumentation framework from Example 7



be an argumentation framework. The corresponding graph can be found in Figure 2. Using Equation 2 we obtain $\Delta = \{(a|\neg b), (b|\neg a), (c|\neg b \wedge \neg a \wedge \neg d), (d|\top)\}$. With $\Delta = \Delta_0$ we have $(\kappa^Z_\Delta)^{-1}(0) = \{abcd, ab\bar{c}d, a\bar{b}cd, a\bar{b}\bar{c}d, \bar{a}bcd, \bar{a}b\bar{c}d\}$. Now we can count the number of occurrences of each argument. So $Ccs^\theta_{\kappa^Z_\Delta(\omega)}(a) = 4$, $Ccs^\theta_{\kappa^Z_\Delta(\omega)}(b) = 4$, $Ccs^\theta_{\kappa^Z_\Delta(\omega)}(c) = 3$ and $Ccs^\theta_{\kappa^Z_\Delta(\omega)}(d) = 6$. This results in $d \succeq^{Ccbs} a \simeq^{Ccbs} b \succeq^{Ccbs} c$. Looking at the graph we see, that argument d is unattacked, so it is intuitive that this argument is ranked at the highest position. Also the arguments a and b are attacking each other and are not attacked by any other argument. These two arguments are there indistinguishable and should be ranked on the same level, but both arguments have at least one attacker so it should be ranking lower then d. Argument c is attacked by three other arguments and defended by none, hence this argument should be ranked lower then its attackers.

For some ideas of other translations we recommend [Heyninck et al., 2020]. Instead of system Z we could also use c-representations [Kern-Isberner, 2001].

Ranking-based semantics are usually evaluated wrt. a series of rationality postulates [Amgoud and Ben-Naim, 2013, Delobelle, 2017]. We want to look at two simple ones and evaluate our semantics with them. Namely *Void Precedence* [Matt and Toni, 2008, Amgoud and Ben-Naim, 2013] and *Self-Contradiction* [Matt and Toni, 2008]. The idea of *Void Precedence* states that a non-attacked argument should be strictly more acceptable than an attacked argument.

Definition 8 (Void Precedence)**.** A ranking-based semantics σ satisfies *Void Precedence* if and only if for any $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ and $\forall a, b \in \mathcal{A}$, if $\forall c \in \mathcal{A}$ $(c, a) \notin \mathcal{R}$ and $\exists d \in \mathcal{A}$ with $(d, b) \in \mathcal{R}$, then $a \succ^\sigma_{AF} b$.

On the contrary, a self-attacking argument should always be ranked worse than any other argument, because these arguments are contradicting themselves. This is handled with the property *Self-Contradiction*.

Definition 9 (Self-Contradiction)**.** A ranking-based semantics σ satisfies *Self-Contradiction* if and only if for any $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ and $\forall a, b\mathcal{A}$, if $(a, a) \notin \mathcal{R}$ and $(b, b) \in \mathcal{R}$ then $a \succ^\sigma_{AF} b$.

Proposition 10. Ccbs *does not satisfy Void Precedence nor Self-Contradiction.*

Hence this semantics has a few shortcomings, but an extension, which solves the problem of self-attacking arguments, should yield a reasonable ranking-based semantics. A full analysis for such a semantics will be presented in a follow-up work.

Another future work approach is to look at other frameworks like ADFs presented in [Brewka et al., 2013], which uses an acceptance function for every argument. This could prove to be helpful in finding a ranking with conditional logic.

[Kern-Isberner and Simari, 2011] used a similar idea to rank arguments from a *Defeasible Logic Programming* (DeLP), a system, which combines logics programming with defeasible argumentation. They used system Z to identify "good" arguments.

4 Conclusion

In this work we have presented a first idea to rank arguments with conditional logics. For this we first looked at a simple transition from an argumentation framework to conditional logic and applied an inference relation. Using a simple counting idea results in a ranking over arguments.

Although this semantics does not satisfy two desired properties, we have a established simple connection between ranking arguments and conditional logic. In the future we can improve this idea and present a ranking-based semantics, which satisfies a good number of properties presented in [Delobelle, 2017].

Acknowledgement

The research reported here was supported by the Deutsche Forschungsgemeinschaft under grant KE 1413/11-1.

References

[Amgoud and Ben-Naim, 2013] Amgoud, L. and Ben-Naim, J. (2013). Ranking-based semantics for argumentation frameworks. In *Proceedings of the 7th International Conference on Scalable Uncertainty Management (SUM'13)*, pages 134–147.

[Brewka et al., 2013] Brewka, G., Ellmauthaler, S., Strass, H., Wallner, J., and Woltran, S. (2013). Abstract dialectical frameworks revisited. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI'13)*, pages 803–809.

[De Finetti, 2017] De Finetti, B. (2017). *Theory of probability: A critical introductory treatment*, volume 6. John Wiley & Sons.

[Delobelle, 2017] Delobelle, J. (2017). *Ranking-based Semantics for Abstract Argumentation*. PhD thesis, Artois University.

[Dung, 1995] Dung, P. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.

[Goldszmidt and Pearl, 1996] Goldszmidt, M. and Pearl, J. (1996). Qualitative probabilities for default reasoning, belief revision, and causal modeling. *Artificial Intelligence*, 84(1-2):57–112.

[Heyninck et al., 2020] Heyninck, J., Kern-Isberner, G., and Thimm, M. (2020). On the correspondence between abstract dialectical frameworks and nonmonotonic conditional logics. In *Proceedings of the 33rd International Florida Artificial Intelligence Research Society Conference (FLAIRS-33)*.

[Kern-Isberner, 2001] Kern-Isberner, G. (2001). *Conditionals in nonmonotonic reasoning and belief revision: considering conditionals as agents*. Springer.

[Kern-Isberner and Simari, 2011] Kern-Isberner, G. and Simari, G. (2011). A default logical semantics for defeasible argumentation. In *Twenty-Fourth International FLAIRS Conference*.

[Kern-Isberner and Thimm, 2018] Kern-Isberner, G. and Thimm, M. (2018). Towards conditional logic semantics for abstract dialectical frameworks. In *Argumentation-based Proofs of Endearment - Essays in Honor of Guillermo R. Simari on the Occasion of his 70th Birthday*. College Publications.

[Matt and Toni, 2008] Matt, P. and Toni, F. (2008). A game-theoretic measure of argument strength for abstract argumentation. In *Proceedings of the 11th European Conference on Logics in Artificial Intelligence*, pages 285–297. Springer.

[Pu et al., 2014] Pu, F., Luo, J., Zhang, Y., and Luo, G. (2014). Argument ranking with categoriser function. In *Proceedings of the 7th International Conference on Knowledge Science, Engineering and Management*, pages 290–301.

[Spohn, 1988] Spohn, W. (1988). Ordinal conditional functions: a dynamic theory of epistemic states. In *Causation in Decision, Belief Change, and Statistics*, pages 105–134. Kluwer.

[Wu et al., 2010] Wu, Y., Caminada, M., and Podlaszewski, M. (2010). A labelling-based justification status of arguments. *Studies in Logic*, 3(4):12–29.

Speech acts and enthymemes in argumentation-based dialogues

Andreas Xydis

Department of Informatics, King's College London, UK

Abstract

In logic-based argumentation arguments typically consist of a conclusion deductively and/or defeasibly inferred from some premises. However, in practice, humans do not always present all the elements of their arguments. Alternatively, they assert arguments with an incomplete logical structure called enthymemes. In this paper, we present locutions, witnessed in real-world dialogues, that handle the use of enthymemes during argumentation-based dialogues. Some of these locutions have not been studied in computational models of argumentation-based dialogues and, thus, they enrich existing systems which capture limited scenarios on how a dialogue can unfold. Additionally, we highlight our PhD objectives.

1 Introduction

When computational agents (computer programs that perform actions autonomously) engage with humans, for example to resolve conflicts, find a proof or cooperate to reach a decision, they need to justify their claims if they want to convince the other party that these claims are valid. This is a key concern of argumentation. By exchanging arguments the agents try to draw conclusions regarding the topic of their discussion. This process is known as argumentation-based dialogue. From now on we will be referring to argumentation-based dialogues as dialogues.

In most works which concentrate on dialogues, arguments are dealt in an abstract way, i.e. arguments are depicted as atomic formulas (abstract argumentation). However, we use logic-based argumentation since it examines the internal structure of an argument, i.e. the components used to instantiate it, thus allowing to explore enthymemes, which are more representative of real life dialogues. If we are to provide normative support for human-human debate and enable AIs and humans to jointly reason, we need to investigate how to process enthymemes during dialogues.

Although there are some works on enthymemes, e.g. [Black and Hunter, 2012, Hosseini et al., 2014, Hunter, 2007, Panisson and Bordini, 2017], most of them focus on how to formalise enthymemes when agents share some knowledge and why an agent might decode an enthymeme in a wrong way. Only a few works explore dialogue systems which employ enthymemes, but they only adopt specific types of dialogues, namely inquiry [Black and Hunter, 2008] and information-seeking [Hosseini et al., 2017] dialogues. Consequently, they express limited cases regarding how a dialogue may develop and the different answers that can be given by a participant.

The objective of our research is to develop a dialogue system which allows the use of enthymemes for various dialogue types [Walton and C. W., 1995] and to generate an argumentation framework (referred to as dialogue framework throughout the rest of the paper) from the moves made, which is used to evaluate the acceptance of enthymemes according to some semantics. Moreover, we want to investigate if there is a correspondence between the dialogue framework created at any stage in the dialogue and the Dung argumentation framework [Dung, 1995] instantiated by the contents of all the moves made at that stage in the dialogue.

In this paper, we present a set of speech acts,

which supports our desired dialogue system, through the use of an example. These locutions are effectively motivated by the everyday use of enthymemes. While some of them have already been used in existing dialogue frameworks, e.g. [Amgoud et al., 2000, Black and Hunter, 2007, Prakken, 2005, Hosseini et al., 2017], there are others that have not been applied in current computational models of dialogues.

2 Method

We formalise arguments within the $ASPIC^+$ framework [Modgil and Prakken, 2013] as, using it, we can apply Dung's theory for evaluating arguments [Dung, 1995], it accounts for the structure of arguments and it can accommodate existing argumentation formalisms, such as deductive argumentation [Bondarenko et al., 1997] and ABA [Besnard and Hunter, 2008]. At the same time it, also, provides guidelines for one to define their own way of constructing arguments into a given logic. This allows us to use, also, the concept of argument-trees [Hosseini et al., 2014] which is necessary to define enthymemes structurally and is compatible with the rest of the principles of $ASPIC^+$.

To instantiate locutions motivated from real life scenarios, we decided to create an example based on a real world law case named Frazier v. Cupp (1969). As a case of a persuasion dialogue, analyzing it gave us useful insights since we are only aware of how enthymemes are handled in inquiry and information seeking dialogues. Afterwards, we formalised a basic dialogue system which provided us with the foundations on how to model an interaction between two agents.

3 Discussion

In this section, we show how locutions found work. Solid arrows depict strict inference rules, meaning that if its antecedents are accepted, its consequent must be accepted. Double line arrows represent defeasible inference rules, meaning that if its antecedents are accepted, its consequent must be accepted unless there are good reasons not to.

Example 1. *The police arrests Frazier with his cousin and tries to prove to Frazier that he is guilty for the death of a person, whereas Frazier tries to persuade the police that he is innocent. The policeman* ***asserts*** *the argument A:*

"You were in the bar where the victim was last seen alive."(premise a) "If you were in the bar where the victim was last seen alive, then you are a suspect."(inference rule $a \rightarrow b$) "Therefore you are a suspect."(conclusion b).

Then Frazier asserts the enthymeme B_1:

"I was not in the bar where the victim was last seen alive."($\neg a$)

based on the argument B:

"I was at my cousin's block of flats the whole time."(c) "My cousin can confirm I was at my cousin's block of flats the whole time."(e) "If I was at my cousin's block of flats the whole time and my cousin can confirm I was at my cousin's block of flats the whole time, then I was not in the bar where the victim was last seen alive."($c, e \rightarrow \neg a$) "Therefore, I was not in the bar where the victim was last seen alive."($\neg a$)

A dialogue must start by an *assert* move in order for the agent to set the topic of the dialogue. Later, using *assert*, an agent moves an argument or an enthymeme to attack a previous argument or an enthymeme in the dialogue. As we can see in Example 1, B_1 is moved against the premise of A since B_1 is the negation of the premise of A. However, Frazier does not give a support for his claim and so the policeman, who is not aware of the argument B, can ask Frazier to justify it. To capture this we introduce the locution **why** which requests from the other participant to provide support for their assertion. Additionally, we propose the locution **because** as an answer to the locution *why*. This speech act allows the participant to either backward expand on their previous enthymeme, i.e. move an argument/enthymeme which consists of the support for their last claim and their claim, or repeat their assertion if there is no support for it. Frazier, now, can use *because* to move B and fulfil the request of the policeman or he can use the locution **stop** to show that he wants to end

the dialogue. An agent can use *stop* as a reply to every speech act that will follow, unless specified.

Example 2. *Continuing Example 1, suppose that Frazier answers to the why locution by using because*(B)*. The policeman, then, asserts the enthymeme* C_1*:*

"Your cousin confessed."(f)

based on the argument C*:*

"Your cousin confessed."(f) *"Your cousin told the police she was asleep."*(t) *"If your cousin confessed and your cousin told the police she was asleep, then apparently she cannot confirm that you were in your cousin's block of flats the whole time."*$(f, t \Rightarrow \neg e)$ *"Therefore, apparently, your cousin cannot confirm that you were in your cousin's block of flats the whole time."*$(\neg e)$

As we can see in Example 2, it is not clear to Frazier why the policeman moved C_1 against B as it does not attack any one of its components and Frazier is not aware of C. In this case Frazier could, naturally, ask the policeman what is implied by the enthymeme that he used, so that Frazier understands the attacking relationship between the two claims. To capture this we introduce the locution **and-so**. This question indicates that the sender of the previous move must forward expand on what they asserted, i.e. reveal what they mean by their enthymeme. To fulfil Frazier's request, the policeman must use the locution **hence** which is used when a participant of a dialogue needs to elaborate on the enthymeme they moved by declaring their syllogism and the conclusion to which it leads. The agent can also repeat their claim, using *hence*, in case there is nothing more they meant by what they revealed. Specifically, in Example 2, the policeman would have to reveal C.

Example 3. *Continuing Example 2, suppose that instead of asking* $and - so$*, Frazier (based on* C_1 *and his own knowledge) assumes that the policeman's intended argument is* D*:*

"Your cousin confessed."(f) *"Your cousin told the police you left her flat for a few minutes."*(u) *"If your cousin confessed and your cousin told the police you left her flat for a few minutes, then apparently you were not at your cousin's block of flats the whole time."*$(f, u \Rightarrow \neg c)$ *"Therefore, apparently, your were not at your cousin's block of flats the whole time."*$(\neg c)$

and so he replies with the argument E*:*

'I was having a phone call at the corridor of my cousin's block of flats."(z) *"The corridor of my cousin's block of flats is part of my cousin's block of flats."*(v) *"If I was having a phone call at the corridor of my cousin's block of flats and the corridor of my cousin's block of flats is part of my cousin's block of flats, then I was at my cousin's block of flats the whole time."*$(z, v \rightarrow c)$ *"Therefore I was at my cousin's block of flats the whole time."*(c)

Let us assume that the policeman is not aware of D. Then he cannot understand why Frazier moved E as a reply to C_1. As before, the policeman can use the locution $and-so$ to ask Frazier what he means by his claim. However, Frazier does not mean anything more than what he revealed, i.e. there are no other conclusions to be drawn based on what he moved. So he answers with the locution *hence* by moving again E. Since Frazier repeated his argument, meaning that there is nothing more to add, the policeman can explore if Frazier understood something else than what the policeman meant by his enthymeme. To capture this, we introduce the locution **what-did-you-understand**. An agent uses this speech act to ask their counterpart what they assumed when they received the agent's argument/enthymeme.

To answer to the $what - did - you - understand$ question, an agent uses the locution **assumed**. With this locution the agent reveals the argument that he believed his counterpart meant by the last argument/enthymeme he moved. Continuing Example 3, if the policeman used the locution $what-did-you-understand$, Frazier would need to use either *stop*, to show that he wants to end the dialogue, or *assumed*, to reveal D and, in this way, explain what he understood based on C_1 and why he moved E.

Suppose that the speech act *assumed* is used by Frazier. Now the policeman has two options: either to correct Frazier and reveal the argument that the policeman really intended or to confirm that what Frazier revealed, using *assumed*, is actually the intended argument of the policeman. To capture this

we introduce the locutions **meant** and **agree**, respectively. In our example, if the policeman chose to use the locution *meant* he would have to reveal C, whereas if he chose the locution *agree* he would have to move D again. Since we have assumed that the policeman was not aware of D, the policeman would have to choose $meant(C)$.

We have gathered all the speech acts explained above, together with the permitted replies for each one, in Table 1. If we compare our proposed locutions with those from other works in dialogues which capture enthymemes, we can see that we have added some new ones ($and - so$, *hence*, $what - did - you - understand$, *assumed*, *meant*). With these new speech acts, firstly, we are able to deal with enthymemes which we can forward expand, i.e. based on the enthymeme moved we can ask for and reveal the conclusion intended by it. Secondly, we can address situations of misunderstandings where the recipient of an enthymeme constructs a different argument than the one intended by the sender.

4 Conclusion

In this paper we introduced locutions inspired from real world dialogues between human agents. Using these locutions we are able to define a dialogue system which handles enthymemes and subsumes a variety of dialogue types. Due to lack of space we were unable to explain other aspects of the developed dialogue system, i.e. its protocol and the dialogue framework generated based on the moves exchanged. Using our speech acts we are able to manage situations in dialogues where we need to backward expand enthymemes (like other works do), forward expand them and clarify any misunderstandings between the participants due to enthymemes. In the best of our knowledge, the last two cases are not integrated by relevant works. We conjecture that if participants play 'logically perfectly' (see [Prakken, 2005]), then the status of enthymemes in the dialogue framework at any stage in a dialogue will correspond with the status of these enthymemes in the Dung argumentation framework instantiated by the contents of all the moves made at that stage in the dialogue. The next step of our research is to explore this assumption. If time permits, we will also investigate how enthymemes can be used for strategic purposes in persuasion dialogues.

Acknowledgements

The research described in this paper could not have been possible without the help of E. Black, C. Hampson and S. Modgil.

Locutions	Replies
assert	*assert* *why* $and - so$ *stop*
why	*because* *stop*
because	*assert* *why* $and - so$ *stop*
$and - so$	*hence* *stop*
hence	*assert* *why* $and - so$ $what - did - you - understand$ *stop*
$what - did - you - understand$	*assumed* *stop*
assumed	*meant* *agree*
meant	*assert* *why* $and - so$ *stop*
agree	*assert* *stop*
stop	*assert* *stop*

Table 1: Locutions and their replies in a dialogue

References

[Amgoud et al., 2000] Amgoud, L., Maudet, N., and Parsons, S. (2000). Modelling dialogues using argumentation. In *Proceedings Fourth International Conference on MultiAgent Systems*, pages 31–38. IEEE.

[Besnard and Hunter, 2008] Besnard, P. and Hunter, A. (2008). *Elements of argumentation*, volume 47. MIT press Cambridge.

[Black and Hunter, 2007] Black, E. and Hunter, A. (2007). A generative inquiry dialogue system. In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent systems*, pages 1–8.

[Black and Hunter, 2008] Black, E. and Hunter, A. (2008). Using enthymemes in an inquiry dialogue system. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent systems*, pages 437–444.

[Black and Hunter, 2012] Black, E. and Hunter, A. (2012). A relevance-theoretic framework for constructing and deconstructing enthymemes. *Journal of Logic and Computation*, 22(1):55–78.

[Bondarenko et al., 1997] Bondarenko, A., Dung, P. M., Kowalski, R. A., and Toni, F. (1997). An abstract, argumentation-theoretic approach to default reasoning. *Artificial intelligence*, 93(1-2):63–101.

[Dung, 1995] Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.

[Hosseini et al., 2014] Hosseini, S. A., Modgil, S., and Rodrigues, O. (2014). Enthymeme construction in dialogues using shared knowledge. In *Proceedings of Computational Models of Argument*, pages 325–332.

[Hosseini et al., 2017] Hosseini, S. A., Modgil, S., and Rodrigues, O. (2017). Dialogues incorporating enthymemes and modelling of other agents' beliefs.

[Hunter, 2007] Hunter, A. (2007). Real arguments are approximate arguments. In *AAAI*, volume 7, pages 66–71.

[Modgil and Prakken, 2013] Modgil, S. and Prakken, H. (2013). A general account of argumentation with preferences. *Artificial Intelligence*, 195:361–397.

[Panisson and Bordini, 2017] Panisson, A. R. and Bordini, R. H. (2017). Uttering only what is needed: Enthymemes in multi-agent systems. pages 1670–1672.

[Prakken, 2005] Prakken, H. (2005). Coherence and flexibility in dialogue games for argumentation. *Journal of logic and computation*, 15(6):1009–1040.

[Walton and C. W., 1995] Walton, D. and C. W., Krabbe, E. (1995). *"Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning"*. State University of New York Press.